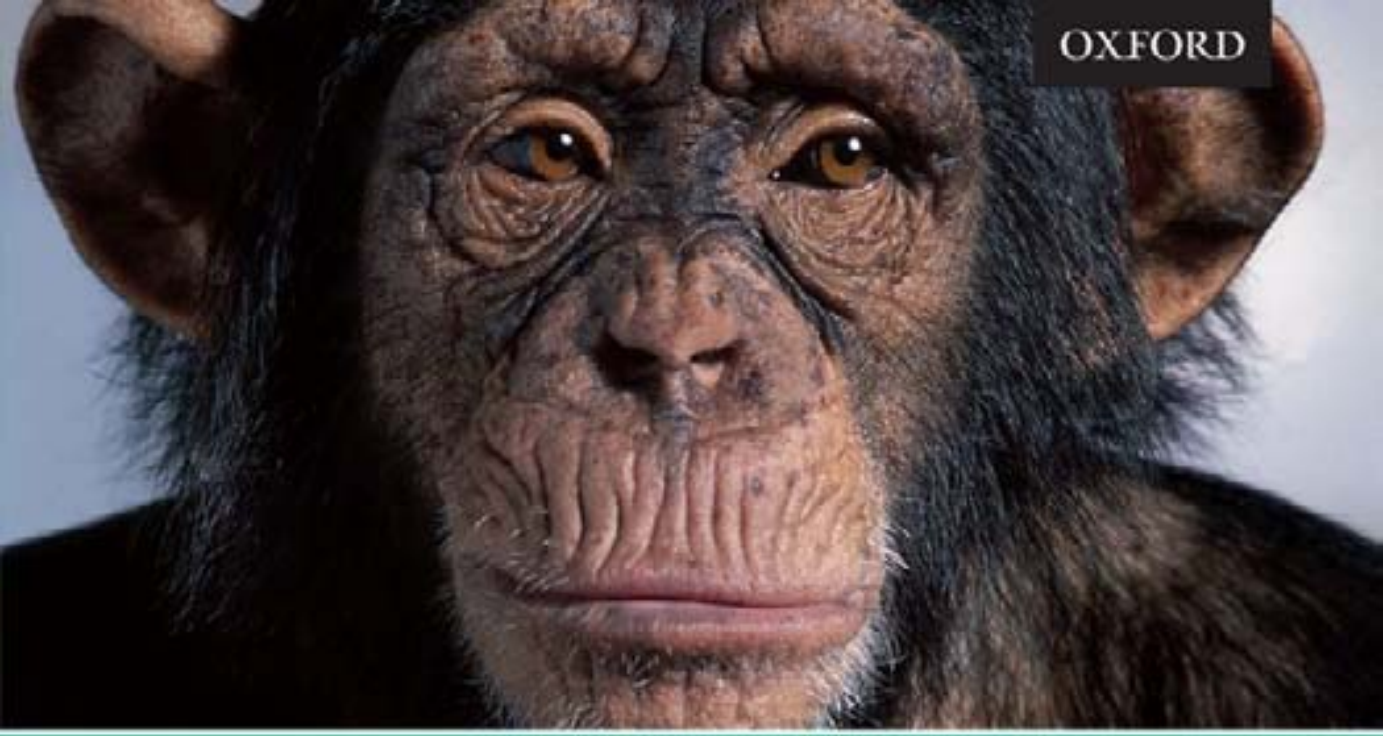


OXFORD

FOUNDATIONS OF METACOGNITION

Edited by

MICHAEL J. BERAN, JOHANNES L. BRANDL,
JOSEF PERNER, JOËLLE PROUST

Foundations of Metacognition

This page intentionally left blank

Foundations of Metacognition

Edited by

Michael J. Beran

Johannes L. Brandl

Josef Perner

Joëlle Proust

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© Oxford University Press 2012

The moral rights of the authors have been asserted

First Edition published in 2012
Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Library of Congress Control Number: 2012944043

ISBN 978–0–19–964673–9

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Oxford University Press makes no representation, express or implied, that the
drug dosages in this book are correct. Readers must therefore always check
the product information and clinical procedures with the most up-to-date
published product information and data sheets provided by the manufacturers
and the most recent codes of conduct and safety regulations. The authors and
the publishers do not accept responsibility or legal liability for any errors in the
text or for the misuse or misapplication of material in this work. Except where
otherwise stated, drug dosages and recommendations are for the non-pregnant
adult who is not breast-feeding

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

Preface

This volume is based on the joint effort of an interdisciplinary group of biologists, psychologists, and philosophers to get a better understanding of the foundations of metacognition. The group worked together for 3 years in an interdisciplinary collaborative research project directed by Joëlle Proust. The group consisted of Michael J. Beran, Johannes L. Brandl, Josep Call, Hannes Leitgeb, Josef Perner, Bernard Renault, and J. David Smith, as well as the collaborators Markus Aichhorn, Alexandre Billon, Mehmet Cakmak, Simone Duca, Frank Esken, Monika Hildenbrand, Daniela Kloo, Anna Loussouarn, Christine Maria Mauer, Bibiane Rendl, Michael Rohwer, and Benjamin Weiss. The project entitled 'Metacognition as a Precursor to Self-Consciousness: Evolution, Development and Epistemology' was part of the Eurocores Programme 'Consciousness in a Natural and Cultural Context' funded by the European Science Foundation as well as by the National Research Agencies of Austria, France, the United Kingdom, and the United States. The editors would like to thank these organizations for their financial support. We also express our gratitude to Eva Hoogland and Chloe Kembery from the European Science Foundation and to Martin Baum and Charlotte Green from Oxford University Press that made this collaboration both pleasant and fruitful. Most of all, however, we would like to thank the authors who accepted our invitation to join this project for their exciting contributions.

Michael J. Beran
Johannes L. Brandl
Josef Perner
Joëlle Proust

This page intentionally left blank

Contents

Contributors

S. R. Beck
University of Birmingham, UK

Michael J. Beran
Georgia State University, USA

Denis Bonnay
Institut Jean-Nicod, Paris

Joseph Boomer
University at Buffalo, The State University of New York, USA

Johannes L. Brandl
University of Salzburg, Austria

Josep Call
Max Planck Institute for Evolutionary Anthropology, Germany

Peter Carruthers
University of Maryland, College Park, USA

Fabrice Clément
University of Lausanne, Switzerland

Kathleen H. Corriveau
Harvard University, USA

Justin J. Couchman
University at Buffalo, The State University of New York, USA

Mariana V. C. Coutinho
University at Buffalo, The State University of New York, USA

Jonathon D. Crystal
Indiana University, USA

Zoltán Dienes
University of Sussex, UK

Jérôme Dokic
EHESS, Institut Jean-Nicod, France

Paul Egré
Institut Jean-Nicod, France

Frank Esken
University of Osnabrück/University of Bielefeld, Germany

Kazuo Fujita
Kyoto University, Japan

Maria Fusaro
University of California at Davis, USA

Paul L. Harris
Harvard University, USA

Sumie Iwasaki
Kyoto University, Japan

Daniela Kloo
University of Salzburg, Austria

Melissa Koenig
University of Minnesota, USA

Asher Koriat
University of Haifa, Israel

Susanne Kristen
University of Munich, Germany

Hannes Leitgeb
University of Munich, Germany

Janet Metcalfe
Columbia University, USA

Noriyuki Nakamura
Chiba University, Japan

Elisabeth S. Pasquini
Harvard University, USA

Josef Perner
University of Salzburg, Austria

Hannah Perst
University of Munich, Germany

Joëlle Proust
Institut Jean-Nicod, France

J. Brendan Ritchie
University of Maryland, College Park, USA

E. J. Robinson
University of Warwick, UK

Michael Rohwer
University of Salzburg, Austria

M. G. Rowley
Keele University, UK

J. David Smith
University at Buffalo, The State University of New York, USA

Beate Sodian
University of Munich, Germany

Lisa K. Son
Barnard College, USA

Claudia Thoermer
University of Munich, Germany

Tillmann Vierkant
University of Edinburgh, UK

Sota Watanabe
Kyoto University, Japan

On the nature, evolution, development, and epistemology of metacognition: introductory thoughts

Michael J. Beran, Johannes L. Brandl, Josef Perner, and Joëlle Proust

The very idea of publishing another book on metacognition needs a word of justification as there is already a number of collections available in this rapidly growing field.[1] The present volume differs from these publications in several important ways.

First, it is to our knowledge the first publication that explores metacognition from a genuinely interdisciplinary viewpoint. Each of the sections in the volume offers various disciplinary angles and methodologies: philosophy of mind, formal semantics, and epistemology are aiming to address the questions initially raised within the comparative, developmental, and experimental psychology of metacognition. Among the questions of common interest are the validity of behavioural methods that test metacognition in non-humans, the reasons for defending a procedural form of metacognition, and the operational definition that could be used to guide investigations into it.

Second, this book addresses questions that are new, or only marginally evoked in other books: in which ways exactly does metacognition relate to metarepresentation and theory of mind? How did metacognition evolve into an explicit mentalizing ability? What is the role of language in this evolution, and in the development of metacognition in children? What kinds of metacognitive abilities are involved when children monitor the reliability of informants? Why might indicative conditionals qualify as instances of procedural metacognition? What kind of semantic content do noetic feelings have, if any? Is the epistemological contrast between uncertainty and ignorance relevant to metacognitive studies?

Third, one of the aims of this volume is to explore the full scope of metacognition. Just as cognition encompasses much more than purely epistemic processes, metacognition applies not only to reflexive epistemic states—knowing that one knows or whether one knows—but to all the

[1] The existing collections on metacognition include: F. E. Weinert and R. H. Kluwe (Eds.) *Metacognition, Motivation, and Understanding* (1987); T.O. Nelson (Ed.) *Metacognition, Core Readings* (1992); J. Metcalfe and A. P. Shimamura (Eds.) *Metacognition: Knowing about Knowing* (1994); L. M. Reder (Ed.) *Implicit Memory and Metacognition* (1996); M. Izaute, P. Chambres, and P. J. Marescaux (Eds.) *Metacognition: Process, Function and Use* (2002); D. T. Levin (Ed.) *Thinking and Seeing. Visual Metacognition in Adults and Children* (2004); H. S. Terrace and J. Metcalfe (Eds.) *The Missing Link in Cognition. Origins of Self-Reflective Consciousness* (2005); J. Dunlosky and R. A. Bjork (Eds.) *Handbook of Metamemory and Memory* (2008); J. Dunlosky and J. Metcalfe (Eds.) *Metacognition* (2009); A. Eflklides and P. Misalidi (Eds.) *Trends and Prospects in Metacognition Research* (2010); and M. T. Cox, A. Raja, and E. Horvitz (Eds.) *Metareasoning. Thinking about Thinking* (2011).

processes that control and monitor the various cognitive functions. The scope of metacognition is then not confined to metamemory and learning-related reflexive states, but extends to perception, motivation, emotion, and arousal. (Whether the control of action belongs to the metacognitive domain remains an interesting subject of controversy.)

Fourth, this volume is unusual in allowing the readers to explore the philosophical dimensions of the concept of metacognition, which are frequently overlooked by philosophers themselves. Given the key role of metacognition in the processes of rational thinking (evaluation and revision of beliefs and intentions), in mental agency, in conscious awareness, and in self-evaluative emotions, it should become a central subject of interest for philosophers from very different specialities, from philosophy of mind, of language, and epistemology to logic and decision theory.

Why is this volume called *Foundations of Metacognition*? Given that the term 'metacognition' has acquired several different meanings in the literature, a general definition of this term is no longer feasible. This introduction explains, first, some of the background assumptions that drive the various usages of this term, and considers what implications these different meanings of 'metacognition' have for drawing the line between cognitive and metacognitive processes. It then offers a presentation of the chapters belonging respectively to the three sections in the volume: metacognition in non-human animals, metacognition in human development, and the functions of metacognition.

Metacognition: background assumptions

The basic questions

We consider the foundational questions about metacognition to comprise the following three main topics:

1. What is metacognition?
2. What are the best methods for studying metacognition in non-human animals and in human development?
3. What role do metacognitive processes play in the overall functioning of a cognitive agent?

We do not assume that these questions need to be addressed in this very order. In fact, our project started with the methodological problems that arise when one studies metacognition at the level of non-linguistic behaviour. Over the past decade, experimental work with animals including macaques, dolphins, and birds made these problems vivid (see Smith 2009). In these experiments, animals are tested for a specific behavioural response that allows them to improve their success rate. For instance, in a discrimination task where the difficulty varies from trial to trial, they can avoid guessing by choosing a so-called 'opt-out' response. As it turns out, animals learn to use this option in a selective way and thus behave quite like a human person would do when reflecting on its own ignorance. But does this similarity in behaviour show that similar cognitive capacities are used by animals and by humans? That has become the bone of contention in the debate about the uncertainty-monitoring paradigm. Whether this methodology successfully uncovers metacognitive abilities, forms the core question of Section I.

Once this question is put on the table, two large avenues open up for further investigation. One way leads into developmental psychology. Thirty years ago, John Flavell called the study of metacognition a 'new area of cognitive-developmental research' (Flavell 1979). Although there is now a large field under the heading of 'metacognition', the question when children begin to know that they know is still largely open. Usually, this question has been addressed with the same methods as assessing knowledge of *other* agents (see Papaleountiou-Louca 2008). But is the assumption here correct that employing a theory of mind is a precondition for metacognition in human

children? Or could metacognition in turn be a precondition for acquiring the social ability to 'read' other persons' minds? These questions are tackled in Section II.

The second avenue leads to the conceptual problem of how to define a metacognitive function. Taken in one sense, monitoring one's own certainty should clearly count as a metacognitive process. That still leaves it open, however, whether an uncertainty response in experiments with animals is also based on such monitoring. One could sidestep this problem by simply taking the notion of 'monitoring' in a much broader sense that also applies to the monitoring of external circumstances causing uncertainty. Then the empirical data would only show that animals, too, monitor uncertainty-producing circumstances, but not that they manifest a metacognitive capacity. The critical point here concerns the relation between cognition, metacognition, and representation. If a cognitive process serves a metacognitive function, does this imply a higher-order representation? Or can the mind monitor and evaluate its own functioning also without employing any metarepresentational capacities? These questions arise in all three sections, but are explicitly addressed in Section III.

The contributors to this volume offer a variety of different answers to these foundational questions. Our goal was not to select papers in support of one particular view on these matters. Rather, we wanted to document the full spectrum of positions that one finds in this rapidly growing field of research. We believe that it is still too early to judge which view will prove to be most fruitful in the future. So we decided to provide a kind of 'road map' that shows where things stand now and what the next steps might be. We took this attitude also with respect to the question 'What is metacognition?'. One certainly would like to have a precise definition of this ambiguous term to begin with. However, providing such a definition is part of the dispute about how to draw the line between simple minds that are merely cognitive and more complex minds that are metacognitive. Different views on the foundations of metacognition therefore go hand in hand with different definitions, or at least with different interpretations of what appears to be the same definition. Since we did not want to constrain the debate by prescribing our own conceptions, we had to pay a price. Contributors to this volume had to explain for themselves how they understand this notion. As a result, different usages of this term are to be found throughout this volume.

What is metacognition?

From its very inception, metacognition has been a 'many-headed monster' (see Brown 1987, p. 105). We will abstain here from offering a full account of the various meanings this term has acquired. Instead we want to use this general introduction to propose a framework for sorting out the different intuitions people have about the content and the extension of this term. In this way, we hope to justify our decision to leave the term 'metacognition' up for grabs.

The conceptual intuitions we want to capture concern both the meaning of the term 'cognition' as well as the prefix 'meta'. The meaning of 'cognition' depends on how this term is used in cognitive science today. However, the usage is not so clear as to determine a precise meaning of this term. Therefore intuitions come into play. To begin with, one might think of a cognitive process as some 'inner process' that helps organisms to adapt their behaviour to external circumstances. But this understanding of a cognitive process would be much too wide unless one adds further constraints. Otherwise, it would also include, for instance, blood circulation as an internal process that furthers a steady body temperature. To rule out such cases, one might restrict the notion of 'inner process' to mental states or events that are accessible to consciousness. That, however, would make the term 'cognitive' too narrow since it would exclude all subpersonal processes that are inaccessible to consciousness. The way out of this problem that is commonly taken is to define cognitive processes as inner processes with a representational function. Cognition takes place when mental representations are formed in order to serve the purposes of the organism.

The term 'representation', as it is used here, requires careful treatment. There is an obvious but important distinction to be made between a representation in the sense of an *object* with semantic properties, e.g. a photograph that shows a certain building, and the *fact* that some object x is a representation of y, e.g. the fact that it shows the house in which Mozart was born (see Dretske 1995). This distinction is connected with different explanatory objectives. If the picture has been taken at a certain location, this fact itself is not a representational fact, but it explains why it is a picture of a house one can see at this location. If the picture is 3×4 inches large, this fact may explain why it fits into my pocket, but it has nothing to do with the fact that it shows where Mozart was born. We therefore have to keep track of what we are talking about: (1) an object with certain properties, (2) properties of this object, (3) facts that have nothing to do with its semantic properties, (4) facts that explain the semantic properties of the object, and finally (5) representational facts that obtain because the object instantiates certain semantic properties.

When it comes to cognitive states and processes, the same distinctions need to be observed. But what does 'observing' mean here exactly? Does it mean that one cannot understand, for instance, what a decision process is unless one understands the representational facts that make it a cognitive process? There seems to be room here for distinguishing different levels of understanding. Thus, one might say that understanding a decision process merely requires knowing that it is based on a goal and on information about how to reach this goal. A fuller understanding may also require that one knows that goals can be represented in different ways, and that different beliefs about how to reach a certain goal can lead to different decisions. As helpful as the distinctions suggested by the representational approach are, they also give rise to questions about our understanding of cognitive processes that are not easy to resolve. We will encounter later in this introduction a critical question that has been raised in this context, namely whether understanding representational facts necessarily requires the conceptual ability to articulate these facts in propositional form. While Dretske subscribes to such a conceptual constraint, it may be lifted in order to allow for a certain type of low-level understanding of cognitive processes.

Let us turn now to the meaning of the prefix 'meta' as it is used in the term 'metacognition'. Intuitively, this prefix indicates a complexity in cognitive functioning that goes 'beyond' what happens in simple cases of perception, memory, or reasoning. But what does it mean to go 'beyond' such basic processes? When one explains metacognition as 'knowing that one knows' or as 'thinking about what one is thinking', this suggests a state or process that leads to declarative (conceptual, propositional) knowledge about one's own mental states. A main strand in metacognition research has taken the term 'cognition of cognition' precisely in this sense (Flavell 1979). It is not clear, however, that this is also the idea expressed when metacognition is described as 'monitoring and controlling of cognition' (Nelson 1996). Here, too, a hierarchy of mental states is suggested because monitoring and control requires one process that gets information about another process. This architecture need not give rise to declarative knowledge, however, and may just instantiate a heuristic that can guide a mental activity. Since in each case metacognition implies a distinction between a 'lower-level' and 'higher-level' process, the hierarchy model remains ambiguous.

To resolve this ambiguity, one needs to answer two critical questions:

1. What are the minimal criteria that distinguish a cognitive and a metacognitive process?
2. What relation must obtain between cognitive processes in order to make one of them a metacognitive process?

We now want to specify three different positions that one can hold in answering these questions.

The first option defines a view that we call *full-blooded representationalism*. Full-blooded representationalism assumes that in order to grasp a semantic property or a representational fact one

needs to understand fully what it means for an object *x* to represent *y*. Only such a full understanding of representation allows one to grasp the difference between a cognitive and a metacognitive process as well as the relations of 'monitoring' and 'control' that bind these processes together. Take, for instance, the decision to travel to Salzburg to visit Mozart's birthplace. From a representationalist point of view this is a cognitive process that includes a representation of the goal (seeing the house) and of the means how to reach the goal (buying a ticket). Compare this with a decision to postpone the decision where to spend one's vacation. This decision may be described as a metadecision since it is a decision about when another cognitive process should take place. But it is a metadecision only if the goal is to modify one's own decision process. If the goal were merely to save some money by observing the market first, this would not yet count as a 'metadecision' even though it can change one's decision process. The crucial requirement is that the 'metadecider' has to form the *intention* to postpone her decision and to understand how the new information she is seeking may change her preferences. According to a full-blooded representationalist, this means that she must represent both her preferences and the way in which new information may change them. Metacognition therefore goes hand in hand with metarepresentation and the conceptual constraints that Dretske places on the understanding of representational facts (see Dretske 1999; Dienes and Perner 2001).

The other options in answering the two earlier questions arise from rejecting full-blooded representationalism. This can be done either in a radical way that leads to a *non-representationalist* conception of metacognition, or in a moderate way that makes room for a position that we want to call *moderate representationalism*.

A radically non-representationalist view about metacognition says that the differences and the relations between cognitive and metacognitive processes can all be spelled out in purely causal terms. Metacognitive processes have causal antecedents that differ in specific ways from the causal antecedents of simple cognitive processes, and that explains their different functions. Perhaps this view is just a straw man and no theorist has ever subscribed to it. Nevertheless, for comparative reasons it is important to have this view on the table as well. With the other views, it shares the intuitive idea that basic cognitive processes monitor 'the world', while metacognitive processes monitor these first-level cognitive processes. Since 'the world' also contains the cognitive processes of other agents, however, they have to be excluded by assuming a causal network that connects the metacognitive and the cognitive processes within a single cognitive agent. The metacognitive processes may now be said to go 'beyond' cognitive processes in the following sense: a cognitive process *Q* monitors a cognitive process *P* when it detects via a causal mechanism the occurrence of *P* without using thereby any information about 'the world'. It is a purely *internal* monitor that has access only to information available within the system. In the same vein, a system may control its own mental functioning not by changing its relation to 'the world', e.g. by moving its body in order to get better information. Rather, we should think of this control as a causal feedback that affects *P* as a result of having been monitored by Q. For instance, consider a person who tends to get nervous whenever she has to make a decision. Her nervous condition then indicates within her cognitive system that some decision is waiting to be made (we may assume that only a pending decision causes such a nervous condition in her), while her nervousness may influence her decision process at least by delaying it. From a causal point of view, there is no need to invoke representational terms to explain this phenomenon. The arousal of nervousness is simply the causal effect of a decision process that is then under the influence of this emotion.

For several reasons, it is questionable that such a case should already count as metacognition. Firstly, the nervousness is dysfunctional and does not support a proper cognitive functioning. Secondly, there is no flexibility in the response that allows for control to occur. And thirdly, the person is not aware of the fact that she is nervous about her decision. These objections show why

a purely causal view may not be a real contender in the field. Yet, sometimes when metacognition is generously attributed to non-human animals, one gets the impression that such a radically non-representationalist conception of metacognition might be involved (see Kluwe 1982).

The third option that we want to consider is *moderate representationalism*. It tries to steer a middle course between the two extremes just described. This can be done either within the representationalist framework provided by Dretske (1995, 1999), or by lifting some of the constraints that define a full-blown representationalist view. Taking the first path, one can state the basic idea of this view as follows: a metacognitive process can represent a first-order state, but without representing the fact that this state has a certain representational function. Take again the photograph that shows Mozart's birthplace as an example. Suppose one puts this picture on a grid paper and then takes a second photograph of this arrangement. This second picture will now clearly show the size of the first picture, but it says nothing about its representational function. If one did not know what the first picture shows, one could not learn this fact from the second picture. The point of the analogy is this: there may be mental representations of cognitive states that represent those states but do not represent them *as cognitive states*. In other words, they may represent a state that happens to be a cognitive state but without making the fact that it is a cognitive state explicit by representing this fact as well. For instance, a decision process can be internally represented as a process that takes a certain amount of time. In this way a cognitive system may compare different decision processes and find out which factors delay the decision-making. All this can be done without knowing what these first-order decisions are about, i.e. without representing them as invoking contentful mental states.

Joëlle Proust has been advocating the view that metacognition requires no metarepresentations (see Proust 2007). But one must be careful in interpreting this position. She is not holding the radical view that metacognitive processes do not represent cognitive processes at all. Nor does she want to deny that metacognitive processes give rise to a distinct type of representational fact. Rather, her view requires a broader conception of what counts as a representational fact, which allows that the obtaining of a representational relation can be represented even when one is lacking the concepts for articulating this fact in propositional form. According to this suggestion, a metacognitive process can be 'about' another cognitive process in virtue of using information about this state—*including information about its representational function*—while retrieving this information in non-conceptual form.

The idea of non-conceptual representation has been influential in philosophy mainly in connection with perception. The claim has been that perceptual experiences represent without categorizing such objects by employing concepts. This does not necessarily mean, however, that a moderate representationalist must conceive of metacognition on the model of a quasi-perceptual inner sense. There may be other ways of explaining how metacognitive feelings can represent cognitive states without representing them conceptually *as* cognitive states. This leads us back to the empirical research on feelings of uncertainty, of confidence, of self-trust, etc. The intriguing idea that moderate representationalism contributes to this research is based on the observation that neither monkeys nor 2-year-old children possess the mental concepts needed for metacognition in the full-blown representational sense. Yet they may have metacognitive feelings that are not just feelings caused by an uncertain environment, as well as a heuristic for monitoring and evaluating their own cognitive performance that cannot be explained in non-representational terms.

Those who employ the contrast between procedural and declarative metacognition should find this idea of moderate representationalism congenial to their view. However, they should also be aware of the fact that this idea can be spelled out in different ways: either taking basic forms of metacognition to be restricted to non-semantic properties, or by invoking the idea of non-conceptual representational facts.

Studying metacognition from scratch

If we were living in a perfect world—or at least in a perfect conceptual space—we would know a priori how the notion of metacognition should be defined. The present volume has been organized on the presumption that we lack such a priori insights. But that should not hinder the research in this area. There is no need to decide in advance what metacognition is before studying its scope and function. Therefore the contributions to this volume can provide new insights into the foundations of metacognition even though they do not reach a consensus on what this fascinating and puzzling faculty is.

Organization of the volume

Metacognition in non-human animals

The first section in this volume is devoted to metacognition in non-human animals. Metacognition is often defined as 'knowing what one does (or does not) know'. Adult humans clearly have feelings of confidence and doubt, and we comment on those feelings. We experience the sense of knowing and not knowing, of remembering and not remembering things. We can respond appropriately to these feelings by reflecting, rethinking, and seeking information before we make decisions, and such responses typically improve the outcomes of our choices. It is these responses that ground the literature on uncertainty monitoring and metacognition. Researchers take human metacognition to indicate important aspects of mind, including hierarchical cognitive control, self-awareness, and consciousness. Thus, metacognition is acknowledged to be one of the most sophisticated cognitive capacities of humans. The question is whether this is a uniquely human capacity, or whether one might see the beginnings of human metacognition in other animals (Smith et al. 2003; Smith 2009). Research into animal cognition has revealed increasingly complex behaviour on the part of many non-human species that may approximate or even match the cognitive abilities of humans (e.g. Smith et al. 1997). Important questions remain about whether non-human animals can demonstrate metacognition at all (e.g. Carruthers 2008; Hampton 2009; Smith 2009). This section presents the most recent evidence offered in support of animal metacognition along with critiques of that evidence, and it is an important debate. Uncovering the phylogenetic roots of metacognition is an important task. The question of whether animals share some aspects of humans' metacognitive capacity will impact the study of animal consciousness and issues relating to the emergence of theory of mind. Research with animals also has the practical value that it can help sharpen theoretical constructs such as uncertainty monitoring in humans, particularly when measured with non-verbal or preverbal populations given the need to devise tests with animals that are language independent. More broadly, learning whether animals are (or are not) metacognitive has implications for understanding how or why conscious cognitive regulation is such a crucial aspect of humans' cognitive system, whereas studying human metacognition in isolation precludes seeing this important ability in the proper evolutionary context.

Developed as the first test of animal metacognition (Smith et al. 1995), the uncertainty-monitoring paradigm has been used with a variety of species and has been modified over time to address concerns about associative mechanisms that might have supported early results with animals (see Crystal and Foote 2009; Jozefowiez et al. 2009; Smith et al. 2009). In uncertainty monitoring tests, animals are given discriminations or memory tests where trial-by-trial difficulty varies. Animals are also given the means to avoid doing any trials of their choosing. This so-called *uncertainty response* has been interpreted as the means by which the animal demonstrates metacognition by knowing (or not knowing) that it will correctly complete a trial through judiciously using the uncertainty response on exactly those trials for which errors are most likely to occur. Early tests using this method produced

intriguing results, but also were open to a number of criticisms regarding the role of associative learning in task performance (e.g. Smith et al. 2008). Couchman, Beran, Coutinho, Boomer, and Smith (Chapter 1) argue that second-generation uncertainty monitoring tasks provide evidence that animals do monitor their own cognitive states and respond adaptively to uncertainty that is experienced in perceptual judgements, memory tests, and conceptual/relational tests. They review recent results that show that some monkeys will use uncertainty responses even when trial-by-trial feedback is made opaque so that reward is not easily associated with specific responses and when tests are designed to be more naturalistic. They conclude that, taken as a whole, the evidence strongly indicates that some animals are metacognitive and can access and use information about their own mental states. Crystal (Chapter 2) disputes this conclusion after presenting an overview of some paradigms used to test animals, and then suggests that these tests still fall prey to concerns that animals may be using basic learning mechanisms to perform these tasks. He also notes that there remain a number of conflicting definitions of what metacognition is, and what constitutes evidence of metacognition. He is correct in this assessment, but it is encouraging that researchers working on the issue of animal metacognition have shown considerable restraint in both trying to show that animals might be metacognitive and, at the same time, policing themselves by carefully considering alternative explanations. So, all is not lost, and here Crystal suggests that researchers could use simulations from computational models of proposed psychological processes (i.e. metacognition and non-metacognition) as a way to determine what animals might really be doing when they are given these tests. This is an excellent suggestion.

Fujita, Nakamura, Iwasaki, and Watanabe (Chapter 3) present data that may indicate that some birds are metacognitive. Pigeons and bantams had to judge visual targets and search for a particular target. After making a choice, the birds had to indicate their confidence in that response by choosing a higher risk, higher reward option or a lower risk, lower reward option. Birds chose the 'safer' option more frequently after they had just incorrectly selected a target than when they had correctly selected a target. Birds also asked for more 'hints' when faced with stimuli they had to select in a sequence when stimuli were novel than when they were familiar. This could suggest that the birds recognized what information they had and what they did not have, and they adjusted their behaviour accordingly, although the authors noted that these results could be the result of other non-metacognitive processes. Call (Chapter 4) presents a different kind of test of animal metacognition, the information-seeking paradigm. Developed as a more naturalistic test that would require far less training than some of the uncertainty-monitoring paradigms, the basic idea is that animals should seek information when they do not have it, and respond accordingly when they do. Originally tested with chimpanzees and children, Call and his colleagues have found that both groups (and, later, other apes) will reach to where food is hidden when they see that it is hidden, but look before they reach when they have not seen the hiding event. As with the uncertainty-monitoring tests, early versions of this test were open to certain criticisms that animals might be using strategies not reliant on metacognition. In this chapter Call presents the newest evidence to counter those concerns. He also concludes with the same caution expressed by many researchers in this area—that although the evidence appears to support some metacognitive processes in animals, there is much work remaining, and he offers constructive ideas about what future animal metacognition tests might involve.

The final two chapters in this section offer critiques of the animal metacognition work in an effort to more closely align the developmental and philosophical traditions in this area with the recently emerged interest in comparative metacognition. Carruthers and Ritchie (Chapter 5) concede various cognitive processes to animals, but argue that because non-human primates appear incapable of forms of mindreading that require attributions of false belief to other agents,

they cannot be capable of metacognitive monitoring of such states. Key to their position is the contrast between two different accounts of the evolution of metarepresentation. One account claims that metarepresentation emerged to help organisms monitor and control their own mental states. The other account is social in nature, claiming that metarepresentation evolved so that organisms could monitor the mental states of others. These accounts offer different predictions for what kinds of things non-human animals might be able to do with regard to mindreading and metacognitive monitoring. Carruthers and Ritchie defend the social account by arguing that so-called uncertainty responses by animals may reflect non-metarepresentational processes. Perner's (Chapter 6) aim is to find the minimal criterion for meta-abilities in non-linguistic creatures. For Perner, there are two critical questions—(1) does the behaviour supposedly reflective of metacognition depend only on the animal's cognitive ability to be in a particular mental state or does it depend on the animal's metacognitive awareness of being in that state?, and (2) does this behaviour really depend on recognizing being in a particular mental state or could it be dependent on confounded external conditions eliciting these states? He states that it is necessary to distinguish whether animals are simply in a state of ignorance or whether they can actually represent that state. Here, the critical point is whether an uncertain animal is responding only to the uncertainty or is responding to its own reflection on being in that state. He concludes that so far no tasks given to animals can answer these questions satisfactorily for the conclusion that metacognition is evident.

It is important to note that Carruthers and Ritchie, Crystal, and Perner do not argue that animals cannot be metacognitive. Rather, the question is whether the methodology used to test animals is sufficient as a measure of metacognition. This highlights the uniqueness of this collection of papers—each chapter takes a different perspective on many of the same experimental procedures and outcomes. There is no firm line drawn in the sand as to whether animals can or cannot be metacognitive—and that is an appropriate stance. Rather, empirical results are offered by those working with animals, but so too are interpretations of those results from alternate perspectives (e.g. comparative psychology, developmental psychology, philosophy). At debate is whether the evidence is sufficient to conclude that some animals may demonstrate monitoring processes that are relevant to better understanding human metacognition. As such, this section presents the most cutting-edge comparative research in metacognition and the most up-to-date debates about the implications of those experiments for understanding the evolutionary foundations of metacognition.

Metacognition in human development

The second section in the volume is devoted to metacognition in human development. To understand this field it is helpful to clarify how *foundational concerns* about *metacognition* differ from (1) *theory of mind* research and from (2) *practical concerns* about children's metacognitive proficiency.

'Theory of mind' is an umbrella term for investigating the ability to impute mental states to agents. Hence metacognition, often defined as 'knowing that one knows', can be and sometimes is seen as part of theory of mind. As an informal language regulation 'theory of mind' tends to be used for research on understanding other people's minds while the concerns about one's own mind tend to be investigated under the label of metacognition. As some have pointed out (see Proust 2007, Chapter 14, this volume), the difference between having a theory of mind and engaging in metacognition may run much deeper and concern not merely the opposition between understanding other minds (theory of mind = other-directed metacognition) and understanding one's own mind (metacognition = self-directed theory of mind). There may be a crucial difference in the cognitive processes involved in knowing whether and when one knows something, on

the one hand, and in finding out whether and when other people know something. Unlike knowledge about other minds, metacognition may or may not (a central topic of dispute) require a conceptual understanding of the mind. It might get away with a simpler self-evaluative heuristic that is variously described as 'procedural', 'non-conceptual', and 'activity based'. Three chapters in this section draw on this (Esken Chapter 8; Brandl Chapter 9) or a similar (Kloo and Rohwer Chapter 10) distinction in their description of metacognition in young children.

From an empirical point of view, the distinction between theory of mind and metacognition is closely linked with the methodological problems in ascertaining whether a higher-order, i.e. metaprocess, is really involved in the observed behaviour. These issues are still hotly debated in theory of mind research, especially with animals and preverbal infants. However, these methodological problems are even worse for demonstrating metacognition. For instance, very young children help a person, who is looking for an object, by pointing to the object's location when the person does not know where the object is but do not do so otherwise. From this we can infer that the child must understand something about the other person's knowledge. Ongoing controversy focuses on the precise nature of this understanding. Does it involve representations of the other's mental state (Call and Tomasello 2008; Baillargeon et al. 2010) or is it based on a set of 'behaviour rules' (Penn and Povinelli 2007; Perner 2010)? Nevertheless, there must be some higher-order mental process in the animal or infant that captures the other's knowledge state in some way, be it as mental state or as behaviour rule.

In case of self-directed metacognition this higher-order process could also be a non-conceptual self-evaluative process. We leave it open how this process ought to be conceived of. Nevertheless, in the case of theory of mind (other-directed metacognition) its existence can be inferred from responses that systematically vary with the other's mental state. In the case of self-directed metacognition, even when one can show that responses vary with the animal's or person's own mental state, one cannot easily infer the existence of a higher-order process, because it could be the mental state itself that creates the covariation of responses.

Call and Carpenter (2001) showed that 2½-year-old children as well as chimpanzees tried to get information about which location has been baited before committing themselves to a choice of location. This was more likely when they had not seen the hiding than when they had observed it. In analogy to the theory of mind case already mentioned, this demonstrates that the child shows different behaviour when someone knows than when that person is ignorant. However, because in this case that someone is the child herself, the interpretation is more difficult. The difference in the child's behaviour might simply be due to her first-order knowledge state (knowing where the bait is as opposed to being ignorant about it). But this is cognition and not metacognition. For metacognition one would have to establish that her behaviour is based on some higher-order mental state or process that captures her first-order state. The fact that second- and first-order states are both states of the same individual and, therefore, could be causally responsible for the observed behaviour, makes it so much harder to determine whether a higher-order process is at all involved.

This example illustrates the methodological difficulties that the empirically-oriented chapters in this section are grappling with and it gives a sense of how foundational these issues can be.

Our focus on 'foundations of metacognition' differs from much of the usual developmental interest in metacognition, which has been mostly concerned with *practical* issues.

Foundational issues concern whether one is in the right ball park, practical issues are about how one performs there. Much of the early seminal work in theory of mind was foundational: it was concerned with whether animals have a theory of mind at all, when it emerges in children, and how children acquire the basic conceptual framework of such a theory. This is still the field's main preoccupation.

Questions of practice are about how well the theory or its particular concepts can be exercised.[2] Once children have the concept of anger they may still differ in how well they can spot whether someone is angry, or what situations tend to make someone angry. Even adults can fail to apply their conceptual understanding of the importance of visual perspective differences in particular communicative situations (Keysar et al. 2003).

In sharp contrast to theory of mind, metacognition research was strongly focused on questions of practice. The developmental work started with metamemory (Flavell 1979): children's knowledge about their own memory, how much they thought they could later recall, what they knew about how to help themselves to remember more. It was assumed that children already had a concept of memory.[3] Later contact with theory of mind sparked some interest in foundational issues (Sodian and Schneider 1990; Lockl and Schneider 2007). The main field of children's metacognition, which has since blossomed into a large educational enterprise (The European Society for Research on Learning and Instruction, EARLI, sports a whole section—Special Interest Group 16—devoted to metacognition), has, however, taken little notice of foundational issues until recently (e.g. Efklides and Misailidi 2010). This explains why contributors to this field are practically absent from among the authors of this volume.

The shortfall of foundational research is particularly noticeable for human development. It is the stepchild among its relatives as the table indicates. So most of the authors in this section make do by looking at early developing abilities in theory of mind, pretence, or volition and try to extract information about underlying metacognitions from these data. Theories that allow for non-conceptual forms of metacognition become influential here as well. However, within these theories, too, a distinction between foundational and practical issues needs to be observed. On the one hand, we may ask when children acquire the principled ingredients for metacognition, the required concepts or heuristics, as the case may be; on the other hand, we may be mostly interested in how well children can make use of these abilities in concrete situations, either by applying their concepts or be executing a certain metacognitive heuristic.

Amount of research	Foundational	Practical
Theory of mind	☺☺☺☺☺☺	😐😐
Metacognition	☹	☺☺☺☺☺☺

The first chapter in this section by Sodian, Thoermer, Kristen, and Perst (Chapter 7) leads us through the amazing findings on theory of mind in infants and other early cognitive achievements with their potential metacognitive implications. Moving up in age, early insights in children's own knowledge (picked up in detail in Chapters 10 and 11) and their ability to verbally refer to their own mental states are described. The chapter then moves on to epistemic vigilance,

[2] It has to be acknowledged that it is hard to draw a sharp line between foundation and practice as the literature on concept possession shows. Some successful practice is needed to detect concept possession. But how much of it is needed to ensure the child has the concept in question and is not using a closely related concept?

[3] New theories about the nature of metacognition have emerged (e.g. Proust Chapter 14, this volume) that deny the involvement of concepts. The distinction we draw here, nevertheless, persists. Foundational concerns are about when children have the principled ingredients, the required concepts, or non-conceptual experiences, as the case may be. Questions of practice concern how well children can make use of these abilities in real situations.

how children decide which informant to trust (treated in detail in Chapter 12), and ends with the importance of metacognitive linguistic input for metacognitive development.

The following two chapters look for indirect evidence of metacognition in data from the second year of life. Esken (Chapter 8), for instance, looks at the emergence of the social emotions of embarrassment and shame at about 1½–2 years and argues that these emotions can only occur on the basis of some metacognitive reflection on one's behaviour having to meet a standard. Brandl (Chapter 9) provides a theoretical argument that recognizing pretence in others, which emerges around the age of 1½ years, makes children aware of their own pretend intentions and generates a metacognitive feeling of liberty of being able to act without being constrained by reality.

The next two chapters focus, among other topics, on children's insight into their ignorance or uncertainty. Kloo and Rohwer (Chapter 10) look first at children's sense of agency. By 3 years, children are quite proficient at realizing which one of two race cars they are controlling. This ability goes beyond the cognitive processes involved in controlling the cars. It requires some form of reflection on the fact that control is exerted, which the authors call pre-reflective metacognition. Then the authors turn to children's insight into their own ignorance. Three-year-olds are quite good at explicitly saying whether they do or do not know the contents of a box when they have seen the object being put inside as opposed to not having any idea of what could be inside. However, their ability to admit their ignorance breaks down when they have seen a range of objects but don't know which one of them was put inside. Recovery from this case of meta-ignorance (not knowing that they don't know) comes surprisingly late around 6–7 years: a good illustration that even children's explicit verbal responses to questions are not a reliable guide to their metacognitive abilities. Beck, Robinson, and Rowley (Chapter 11) provide further support for such a potentially late understanding by reviewing recent work on children's handling of uncertainty. Although 4–7-year-old children appreciate 'physical' uncertainty about the future (the physical world can turn out one or the other way), they have marked problems with 'epistemic' uncertainty (the world is in a particular state but the child doesn't know yet which) until about 6 years. This suggests that a metacognitive understanding of uncertainty may be in place only at this relatively late age. However, when asked for a confidence judgement of their knowledge, even the 7-year-olds failed to give appropriate confidence judgements of their knowledge.

Finally, Harris, Corriveau, Pasquini, Koenig, Fusaro, and Clément (Chapter 12) look at whose information children trust when given contradictory information. Three-year-olds prefer information from a familiar caregiver over that from a stranger, while 5-year-olds also prefer an informant who has proven more accurate and knowledgeable on a prior test: a tentative indication that by 5 years of age children metacognitively evaluate informants against their own knowledge.

In sum, the chapters show that, although a 'pre-reflective', 'non-conceptual' (as go some of the terms used) metacognition seems to appear very early, full 'reflective', 'conceptual' metacognition does not seem in place until surprisingly late; in the case of knowledge and uncertainty as late as 6 or 7 years or possibly even later.

The functions of metacognition

The goal of the third section is to explore the functions of metacognition, a topic that is clearly at the heart of the controversy reflected in all chapters of this volume. One controversial issue involves *what metacognition was selected for*. A first theory that, from Flavell's pioneering studies (Flavell 1987), has inspired much of the developmental research on children's metacognition, is that metacognition, as a capacity, consists in self-directed mindreading. It is a by-product of a general ability whose function is to enable us to make sense of people's behaviour by ascribing to them mental representations such as beliefs, desires, and intentions. On this view, knowing that

one perceives an object, or evaluating how well one perceives it, requires that one possesses the concept of perception, and is able to apply it correctly to one's own cognitive states as well as to those of others. A second theory, which has fuelled considerable comparative research over the past decade, claims that metacognition is a set of procedures allowing agents to control and monitor their first-order cognitive abilities. On this view, metacognition results from selective pressure, on cognitive systems, to adjust their epistemic goals (such as discriminating, remembering, reasoning) to their cognitive resources. Neither forming the goal of (e.g.) remembering, nor monitoring one's memory, on this view, requires that systems use mental concepts to refer to their own cognitive states. Forming the goal of remembering is practically prompted by a task that requires it. Monitoring one's memory, on the other hand, requires an ability to extract, from current cognitive activity, predictive heuristics about likely success or failure. What distinguishes monitoring a motor from a cognitive activity (e.g. gauging likely success in jumping or in remembering) is, first, that the cues involved in prediction are different (respectively perceptual cues from the environment read by the motor system, and endogenously-generated cues, such as ease of retrieval), and, second, that the norms guiding evaluation are respectively instrumental (such as goal-efficiency and reward) and epistemic (such as accuracy).

On this view, then, systems endowed with metacognition, and sufficiently trained in a cognitive task, can learn how to reliably predict success in this task, even though they do not form conceptual representations about their cognitive capacities.

Several essays in this section discuss the issue of the function and scope of metacognition. Asher Koriat (Chapter 13) recognizes that one type of human metacognition is 'information-based', i.e. it relies on analytic inferences drawing on naïve conceptual theories about mental functioning. People may evaluate their performance on the basis of their beliefs about their own skills and competence, or of the evidence they have for their first-order epistemic judgements. Another type of metacognition, however, is 'experience-based', which means that, in humans, metacognitive judgements can be derived from feelings rather than from theoretical assumptions. Experimental studies on human metacognition in the last three decades have revealed that subjects are able to extract implicit cues such as the fluency with which they select or retrieve an answer to form reliable predictions about the accuracy of their performances. Such activity-dependent, cue-based heuristics are automatically extracted, and accessed only through the conscious experiences they generate, i.e. a more or less intense noetic feeling. In his chapter, Koriat hypothesizes that judgements of self-confidence might in many cases be based on a self-consistency heuristic, i.e. on the proportion of unconsciously sampled representations favouring compatible versus incompatible, alternative answers. Self-consistency being lower for minority than for majority choices in a group, it turns out to correlate with consensuality—independently of the truth of the associated judgements. From this intriguing result, emerges a possible additional function of metacognition as an indirect indicator of consensus—be it for real-world knowledge, or for religious beliefs.

Another way of defending an experience-based form of metacognition consists in showing that systems unable to read minds can still reliably assess the accuracy of their perception or memory in a given task. In Chapter 14, Joëlle Proust reviews comparative evidence supporting the presence of a procedural form of metacognition in non-humans. These findings, along with dissociations, in humans, between epistemic predictions based respectively on experience of a task and conceptual knowledge, suggest that different informational mechanisms are involved, respectively, in attributing mental states to oneself and in evaluating one's own cognitive performances. A double-accumulator model offers an interesting way of abstractly describing the informational mechanisms that might allow a subject to predict her cognitive performances exclusively on the basis of the dynamic properties of the content vehicle (relative to a stored norm). Such a model

has proved able to correlate metaperceptual assessment with the pattern of neural activity in dedicated areas of monkeys' brains.

Hannes Leitgeb approaches the issue of the function of metacognition from a different angle (Chapter 15). Considerations from the formal semantics of indicative conditionals, he claims, offer reasons for defending a procedural, or 'experience-based', level of human metacognition. Testing for the subjective acceptability of an indicative conditional triggers a form of conditional reasoning in which one represents a possible situation, and, based on it, assesses whether this situation entails a given consequence. For example: 'If Oswald did not kill Kennedy, someone else did'. Why should such a form of reasoning qualify as an instance of procedural metacognition? Accepting an indicative conditional, Leitgeb says, is an instance of cognition about cognition: from the supposition that *A*, one comes to believe hypothetically that *B*, and accepting if *A* then *B* expresses this very fact; but this process does not involve a representation of itself as being mental. Lewis' (1976) result in probabilities of conditionals indeed shows that a propositional representation of this kind would be inconsistent with the standard axioms of probability when non-trivial assumptions on the possible subjective probability measures are accepted. Leitgeb thus proposes that an indicative conditional expresses the subject's high subjective probability of *B* given *A* without representing *that* this probability is high (in the same way as a feeling of knowing expresses an epistemic confidence in one's memory without representing *that* one can retrieve an item from one's memory).

In Chapter 17, Tillmann Vierkant offers a critical analysis of the view defended by two philosophers, Philipp Pettit and Victoria McGeer (2002), that one way of explaining the difference between human and non-human minds is that human minds are able to regulate themselves by thinking about content as content, thanks to a linguistic representation. Vierkant agrees that there are forms of self-regulation that do not require representing mental states as states. However, Vierkant has two worries. First, primate evidence shows conclusively that various forms of intentional control of one's cognitive and motivational dispositions do not require linguistic representations. Second, Pettit and McGeer do not take into account that certain forms of self-directed cognitive control require metarepresentations. According to the so-called agency theory of self-knowledge, which Pettit and McGeer defend, what is special about self-knowledge is that, in order to find out what you believe, you merely have to deliberate about the first-order question. If such was always the case, agents would never be wrong in knowing what they believe (and would not need mindreading to know it). There are cases, however, where we are interested in acquiring specific mental states, i.e. attitudes with prespecified contents. In Pascal's wager, for example, one needs to use psychological knowledge in order to make oneself believe in God (in case God exists, a rational goal to have: even if the existence of God cannot be ascertained, a rational being should wager as though God exists, because an immortal life can be gained, while there is nothing to lose). Controlling one's future motivations and goals over time similarly requires that one takes oneself to be a psychological creature, whose states can be manipulated. The author concludes that if self-control is crucial for autonomy, an understanding of folk psychology is crucial for being an autonomous agent.

Whether the function of metacognition is claimed to consist in mindreading or in a specialized form of procedural know-how, another salient question needs to be addressed: how essential is it, for a subject endowed with metacognition, to be consciously aware of having the relevant cognitive states? Does metacognition necessarily engage conscious awareness? The stance taken on this issue obviously depends on what the function of metacognition is taken to be.

Let us assume that metacognition depends on a process of propositional self-attribution. On this view, it is tempting to consider that metacognition coincides with conscious thinking of a first-order proposition. On David Rosenthal's influential view about the nature of consciousness,

a mental state is conscious if one represents oneself as being in that state (Rosenthal 2000a, 2000b). This second-order thought is not itself conscious, unless it is represented as my thought by a third-order thought. Therefore, if one accepts a higher-order theory of consciousness (HOT), granting that metacognition is using metarepresentations entails *ipso facto* recognizing that metacognition makes people conscious of their first-order thoughts. Zoltan Dienes (Chapter 16) accepts both a metarepresentational, theory-based view of metacognition, and Rosenthal's higher-order theory of consciousness. Given these assumptions, a subject who cannot form an accurate higher-order thought about her executive control (and thus has executive but non-HOT—i.e. 'cold'—control over her mental states), is *ipso facto* deprived of metacognition. On this view, hypnosis provides a 'showcase' for metacognition researchers: if the hypnotic suggestion is that the subject's arm will rise by itself, the person raises her arm intentionally, but inhibits her knowledge that she intends to raise her arm. In this case, a perturbed metacognition—understood, here, as the ability to metarepresent one's intentions—offers a cogent explanation of the strange phenomenon of denial of agency in hypnotized subjects. No change in first-order abilities, however, occurs under hypnotical suggestion. Dienes reports that highly hypnotizable subjects appear generally prone to form inaccurate higher-order thoughts; they may thus easily forgo their higher-order thoughts of intending in order to respond hypnotically to suggestions.

In their chapter, Janet Metcalfe and Lisa Son also accept that the function of metacognition is to introspect one's knowledge states, in order to guide cognitive decisions (Chapter 18). Defined in this broad way, metacognition involves some form of consciousness. They argue, however, that metacognitive tasks need not always involve high-level consciousness, and an explicit representation of self. Endel Tulving's tripartition between three forms of consciousness: anoetic (bound to current here and now), noetic (involving semantic memory), and autonoetic (involving a self) is used by Metcalfe and Son to distinguish three kinds of metacognitive judgements (Tulving 2005). Anoetic (meta)cognition includes judgements assessing the value of an external stimulus. This definition, the authors observe, stretches the definition of metacognition to the breaking point, and should rather help reject, as irrelevant to metacognition, a class of tasks currently used in metacognitive studies. Noetic metacognition is a judgement made on the basis of an internal representation of an absent stimulus. Judgements of learning and wagering belong to this category. Autonoetic metacognition, finally, includes judgements that are specifically self-referential, such as source judgements, remember-know judgements, and agency judgements. Thus, the authors conclude, performing autonoetic metacognitive tasks allows agents to be self-aware, while the other types of tasks do not. This contrast, on the view defended, suggests interesting new ways of testing whether non-humans have a self and are aware of it.

The authors who assume that metacognition can develop independently from mindreading tend to recognize a form of consciousness constituted by metacognitive experiences, which does not presuppose the ability to form self-directed higher-order thoughts. They also are more willing to accept that metacognition can be exercised without consciousness. In favour of the latter view, Lynne Reder and colleagues have shown in a series of studies that unconscious strategy selection and monitoring of cognitive performance can occur. Feelings of knowing, in particular, might not always need to be consciously experienced to influence the retrieval process, i.e. to trigger a contextually correct strategy selection (Diana and Reder 2004). In spite of a different terminology, this view is compatible with Asher Koriat's view, that noetic feelings reflect at a conscious level the predictions based on implicit heuristics, whereas information-based, analytic, declarative metacognition is always explicit and, thus, available to conscious report (Koriat 2000).

In Chapter 19, Jérôme Dokic discusses four ways of understanding the nature and epistemic value of noetic feelings. On the 'Simple Model', noetic feelings are manifestations of metarepresented epistemic states. On the 'Direct Access Model', they are partly opaque experiences about

one's own first-order (not-metarepresented) states of knowledge. On the 'Water Diviner Model', they are bodily experiences that are only contingently associated with first-order epistemic states. Finally, the 'Competence' model, which he favours, holds that what appears to be metarepresentational information carried by the intentional content of a noetic feeling has the form 'I can do this' (or the selfless form 'This can be done'), where the demonstrative 'this' refers to the contextually active cognitive task. In this respect, noetic feelings are akin to feelings of physical competence. Although these feelings can be generated by activity-based processes, they are inevitably redescribed in metarepresentational terms by agents endowed with mindreading. Dokic's chapter also emphasizes the difference between feelings being called metacognitive in virtue of their intentional contents (knowledge states) or in virtue of their implicit causal antecedents (their being part of monitoring mechanisms sensitive to fluency and other dynamic aspects of processing). In procedural metacognition, on his view, conscious feelings are only epiphenomenal, because implicit heuristics are sufficient to monitor cognitive activity. Deliberate, i.e. declarative metacognition, in contrast, offers a causal role to feelings in virtue of their intentional contents.

Epistemologists Paul Égré and Denis Bonnay, in the final chapter of this volume, explore the distinction between two states of ignorance, uncertainty and unawareness, a distinction that is of major relevance to metacognitive studies as well as to epistemology, and also complicates further the ways in which metacognition relates to consciousness. Uncertainty has to do with the strength of one's evidence, whereas unawareness is related with a lack of conception, i.e. with a lack of acquaintance with concepts, or with an epistemic failure to entertain a relevant possibility. A main difference between the two states is that unawareness is not spontaneously open to conscious introspection, while uncertainty is. Both uncertainty and unawareness are sources of unknown unknowns and of unknown knowns. In cases of uncertainty, they respectively correspond to overconfidence and underconfidence judgements. In cases of unawareness, unknown unknowns are typical of lack of acquaintance, whereas unknown knowns correspond to cases of failure in detecting relevance among concepts. The authors reason that metacognition, understood as the ability to assess whether one knows a proposition or does not, will have a different pattern, depending on whether it involves checking on the informational content of one's first-order knowledge (typically in how discriminant the information one has turns out to be) or its conceptual content. They illustrate the different metacognitive responses elicited in each case, through a discussion of various experimental studies, and hypothesize, in line with Koriat's chapter, that what makes self-evaluation harder in cases of uncertainty than in cases of unawareness, is that we only need to sample and weight evidence in the first class of cases.

References

Astington, J. W. and Gopnik, A. (1991). *Developing understanding of desire and intention*. In A. Whiten (Ed.) *Natural theories of mind: The evolution, development and simulation of everyday mindreading*, pp. 39–50. Oxford: Basil Blackwell.

Baillargeon, R., Scott, R. M., and He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14(3), 110–18.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In: F. E. Weinert and R. H. Kluwe (Eds.) *Metacognition, Motivation, and Understanding*, pp. 65–116. Hillsdale, NJ: Lawrence Erlbaum.

Call, J. and Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition*, 4, 207–20.

Call, J. and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12, 187–92.

Carruthers, P. (2008). Meta-cognition in animals: a skeptical look. *Mind and Language*, 23, 58–89.

Crystal, J. D. and Foote, A. L. (2009). Metacognition in animals. *Comparative Cognition and Behavior Reviews*, 4, 1–16.

Diana, R. and Reder, L. M. (2004). Visual vs. verbal metacognition: Are they really different? In D. T. Levin (Ed.) *Thinking and Seeing: Visual Metacognition in Adults and Children*, pp. 187–201. Westport, CT: Greenwood/Praeger.

Dienes, Z. and Perner, J. (2001). The metacognitive implications of the implicit-explicit distinction. In P. Chambres, M. Izaute, and P.-J. Marescaux (Eds.) *Metacognition. Process, Function, and Use*, pp. 171–89. Boston, MA: Kluwer Academic Publishers.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: The MIT Press.

Dretske, F. (1999). The mind's awareness of itself. *Philosophical Studies* 95(1–2), 103–24.

Efklides, A. and Misailidi, P. (Eds.) (2010). *Trends and prospects in metacognition research*. New York: Springer.

Flavell, J. H. (1987). Speculation about the nature and development of metacognition. In F. Weinert and R. Kluwe (Eds.) *Metacognition, motivation, and understanding*, pp. 21–9. Hillsdale, NJ: Lawrence Erlbaum.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–11.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comparative Cognition and Behavior Reviews*, 4, 17–28.

Jozefowiez, J., Staddon, J. E. R., and Cerutti, D. (2009). Metacognition in animals: How do we know that they know? *Comparative Cognition and Behavior Reviews*, 4, 29–39.

Kluwe, R. H. (1982). *Cognitive knowledge and executive control*. In: D. R. Griffin (Ed.) *Animal Mind—Human Mind*, pp. 201–24. Berlin: Springer Verlag.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149–71.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, 85, 297–315.

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89(1), 25–41.

Lockl, K., and Schneider, W. (2007). Knowledge about the mind: Links between theory of mind and later metamemory. *Child Development*, 78, 148–67.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, 51, 102–16.

Papaleountiou-Louca, E. (2008). *Metacognition and Theory of Mind*. Newcastle: Cambridge Scholars Publishing.

Penn, D. C. and Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society of London, Series B*, 362, 731–44.

Pettit, P., and McGeer, V. (2002). The self-regulating mind. *Language and Communication*, 22, 281–99.

Perner, J. (2010). Who took the cog out of cognitive science? – Mentalism in an era of anti-cognitivism. In P. A. Frensch, and R. Schwarzer (Eds.) *Cognition and Neuropsychology: International Perspectives on Psychological Science (Volume 1)*, pp. 241–61. London: Psychology Press.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese* 2, 271–95.

Rosenthal, D. (2000a). Consciousness, content, and metacognitive judgments. *Consciousness and Cognition*, 9, 203–14.

Rosenthal, D. (2000b). Metacognition and higher-order thoughts. *Consciousness and Cognition*, 9, 231–42.

Smith, J. D. (2009). The study of animal metacognition. *Trends in Cognitive Sciences*, 13, 389–96.

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R., and Erb, L. (1995). The uncertain response in the bottlenosed dolphin (*Tursiops truncatus*). *Journal of Experimental Psychology: General*, 124, 391–408.

Smith, J. D., Shields, W. E., Schull, J., and Washburn, D. A. (1997). The uncertain response in humans and animals. *Cognition*, 62, 75–97.

Smith, J. D., Shields, W. E., and Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26, 317–73.

Smith J. D., Beran, M. J., Coutinho, M. V. C., and Couchman, J. J. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin and Review*, 15, 679–91.

Smith, J. D., Beran, M. J., Couchman, J. J., Coutinho, M. V. C., and Boomer, J. (2009). Animal metacognition: Problems and prospects. *Comparative Cognition and Behavior Reviews*, 4, 40–53.

Sodian, B. and Schneider, W. (1990). Children's understanding of cognitive cuing: How to manipulate cues to fool a competitor. *Child Development*, 61(3), 697–704.

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? In H. S. Terrace and J. Metcalfe (Eds.) *The Missing Link in Cognition*, pp. 4–56. New York: Oxford University Press.

Section I

Metacognition in non-human animals

This page intentionally left blank

Chapter 1

Evidence for animal metaminds

Justin J. Couchman, Michael J. Beran,
Mariana V.C. Coutinho, Joseph Boomer,
and J. David Smith

Introduction

Humans frequently encounter situations filled with uncertainty. Consider the plight of students taking multiple-choice exams. When they encounter easy questions for which they know the answer, they immediately respond. They know that they know, and use that knowledge to make a quick and confident decision. When they encounter a difficult question, they often recognize their own uncertainty. They know that they do not know, and use that information to engage in different response strategies. For example, they might try to re-think the question or approach it from another angle. They might skip the question and move onto the next. They might ask the instructor for a hint, or seek additional information from the exams of nearby students. All of these strategies are motivated not only by the difficulty of the particular question, but also by the students' recognition and understanding of their own uncertainty. If all goes well, the ability to metacognitively monitor their mental states will allow them to respond adaptively and engage in appropriate strategies that result in higher exam grades.

The study of metacognition, or thinking about thinking, focuses on the self-reflective strategies just described and their impact on decision-making. In human paradigms, these behaviours include hint-seeking, judgements of learning, feelings of knowing, tip-of-the-tongue states, confidence ratings, and the like (Flavell 1979; Nelson 1992; Koriat 1993; Schwartz 1994; Benjamin et al. 1998). Humans can monitor their mental states and change their behaviour to correspond to feelings of confidence or uncertainty (Nelson 1996). This monitor-and-control ability is linked to executive functioning and hierarchical layers of cognition (Nelson and Narens 1990), cognitive self-awareness, and self-regulation (Nelson 1996). We use not only our knowledge, but our knowledge about our knowledge—the strength, reliability, etc.—to make better and more informed decisions. Metacognition also has been linked to mirror self-recognition and the kind of self-awareness it might indicate (Gallup 1982). Because human introspective metacognition often involves explicit awareness, it has been linked to consciousness (Koriat 2007). It is also related to theory of mind and social cognition, because understanding our own thoughts seems prima facie similar to understanding the thoughts of others (Carruthers 2009; Couchman et al. 2009). For these reasons, metacognition is considered a highly sophisticated mental capacity that may be uniquely human (Metcalfe and Kober 2005).

It is thus very important to investigate how humans came to have this ability. Did it gradually evolve from more primitive cognitive abilities, or did it emerge suddenly? Was the evolution of metacognition related to, or even reliant on, other sophisticated mental abilities such as theory of mind, explicit/representational processing, or language? If there is some relation, is metacognition the underlying basis for (some of) these abilities or simply a beneficial side effect? To answer

these questions, we must ask whether animals have anything like human metacognition (Terrace and Metcalfe 2005; Smith 2009). Discovering the existence of animal metaminds—minds capable of thinking about mental states—and determining the relationship between metacognition and other sophisticated abilities those animals might have (or lack) could allow us to map the transitional forms between primitive cognition and human consciousness.

Accordingly, Smith and his colleagues introduced a new area of comparative research by asking whether animals were capable of monitoring their ongoing mental processes (Smith et al. 1995, 1997). Active research continues in this area (Call and Carpenter 2001; Hampton 2001; Basile et al. 2009; Washburn et al. 2009; Couchman et al. 2010; Call Chapter 4, this volume; Fujita et al. Chapter 3, this volume). In this chapter we will present the evidence for animal metacognition as well as some empirical and theoretical challenges to the claim that some animals are metacognitive. Then, we will outline the most recent experiments designed to overcome those challenges and briefly discuss the implications this research has for the evolutionary emergence of the reflective mind.

We believe that, taken as a whole, the evidence indicates that animals possess some metacognitive ability, and that studying this ability will allow us to better understand the emergence of human metacognition. Like the study of mindreading in young children (Clements and Perner 1994; Low 2010), studying subjects on the verge of adult human or human-like cognition can allow us to better understand the evolution of explicit processing, explicit metacognition, metarepresentation, and eventually human consciousness. And, like the mindreading research, it is not always clear exactly where animal metacognition falls on the spectrums of implicit to explicit processing, monitoring to metarepresentation, or cognition to consciousness. We believe that the results discussed in the following sections show that some animals have metaminds, and that further research must be done to determine the full extent and evolutionary implications of their abilities.

Initial evidence for animal metacognition

Most traditional human metacognition paradigms used verbal reports or states (e.g. feelings of knowing, tip-of-the-tongue states). These worked well for humans, because humans often have metacognitive experiences that are subject to verbal description and communication. However, these paradigms are obviously insufficient for testing animals. To overcome this initial problem, early animal paradigms focused on perceptual discriminations that could be designed to incorporate varying levels of difficulty to test the hypothesis that some animals, in addition to perceiving the stimuli, could track their subjective feelings or intuitions about them. These paradigms usually incorporated two primary responses that were objectively correct or incorrect for a given trial. They brought about food rewards or penalty timeouts similar to those in most animal learning or psychophysical tasks. Additionally, animals were given a secondary response option that allowed them to escape the current trial or otherwise avoid having to make that primary response at the present time. In some cases this secondary response—often called the uncertainty response—brought about a less desirable food, a hint, or a guaranteed-win or easy trial. However, as we will describe, rewarding the uncertainty response proved to be theoretically problematic and has largely been abandoned. In most modern paradigms, the uncertainty response simply allows the animal to avoid the current trial and move onto the next. In this sense, it is functionally equivalent to saying 'I don't know', or skipping a difficult question on an exam; it allows a subject to avoid the problem, for better or worse.

This uncertainty response paradigm began with Smith et al.'s (1995) study of a bottlenose dolphin (*Tursiops truncatus*). The dolphin was given a tone discrimination task featuring one

standard high tone (exactly 2100 Hz) and a variety of low tones (ranging from 1200–2099 Hz). Each trial presented the dolphin with a tone and asked the animal to make one of three responses. One primary response lever was associated with the high tone, and was rewarded only when pressed on a trial that presented the high tone. The other primary response lever was associated with the low tones, and was rewarded on trials that presented any tone lower than 2100 Hz. The dolphin also had an optional uncertainty response that opted-out of the current trial and moved onto the next trial. Fig. 1.1a shows that when presented with easy discriminations, the dolphin made many correct primary responses and very few uncertainty responses. As the task became more difficult, uncertainty responding increased and eventually peaked near the animal's perceptual threshold (at about 2086 Hz, just 0.11 semitones from the high tone). Uncertainty responding peaked close to the psychophysical space in which primary responding converged, showing that the dolphin was adaptively declining the most difficult trials.

These behavioural responses were mirrored by ancillary uncertainty behaviours that were thought to be characteristic of human and animal uncertainty (e.g. hesitating, wavering). Fig. 1.1b shows that these behaviours were most frequent precisely when the animal was faced with difficult perceptual-threshold trials. These behaviours seem to suggest that the animal was indeed experiencing uncertainty, while the adaptive opt-out responding suggests that it was monitoring this mental state and adjusting behaviour accordingly. Of course, there are a variety of interesting aspects to these behaviours, most of which have not been explored further. The dolphin might have been using the physical symptoms of uncertainty as a response cue, bypassing the need to directly monitor its mental states. Or, the behaviours might have resulted from uncertainty monitoring and might be indicative of more sophisticated self-control or re-evaluation processes—similar to the experience of the students in the opening example. That these ancillary behaviours have not been more closely examined in other studies is a clear shortcoming of the field, and future research should certainly attempt to address this problem.

The same pattern of using an uncertainty response was found in rhesus monkeys (*Macaca mulatta*) in a computer-based pixel discrimination task (Smith et al. 1997). Monkeys used a joystick to make dense (2950 pixels) or sparse (2949 or fewer pixels) primary responses. They also had the option of using an uncertainty response that brought about a new, guaranteed-win trial. The paradigm was just like Smith et al.'s (1995), with difficulty increasing until the animals' perceptual thresholds were reached. Like the dolphin, monkeys made accurate primary responses when the trials were easy, and made adaptive uncertainty responses when the discriminations were near their perceptual threshold. Interestingly, humans were given the same task and showed an almost identical pattern of uncertainty responding. Furthermore, humans attributed their uncertainty responses to conscious feelings of uncertainty on the most difficult trials.

Given this isomorphic pattern of human and animal responding to uncertainty-monitoring tasks and humans' reports of conscious uncertainty being the motivating factor behind the responses, it may be tempting to conclude that dolphins and rhesus monkeys have metacognition. However, attributing any sophisticated mental ability to animals, especially one so closely linked to declarative consciousness, must be done with caution. A careful consideration of those data indicated that any strong claims for animal metacognition were not yet warranted.

Problems with the initial evidence

As with any claim in comparative psychology, the first question ought to be whether there is some simpler psychological process underlying the behaviour (Morgan 1906). Seemingly complex behaviours are sometimes the result of simple processes. Several theoretical and methodological issues have arisen concerning the nature and interpretation of the uncertainty response

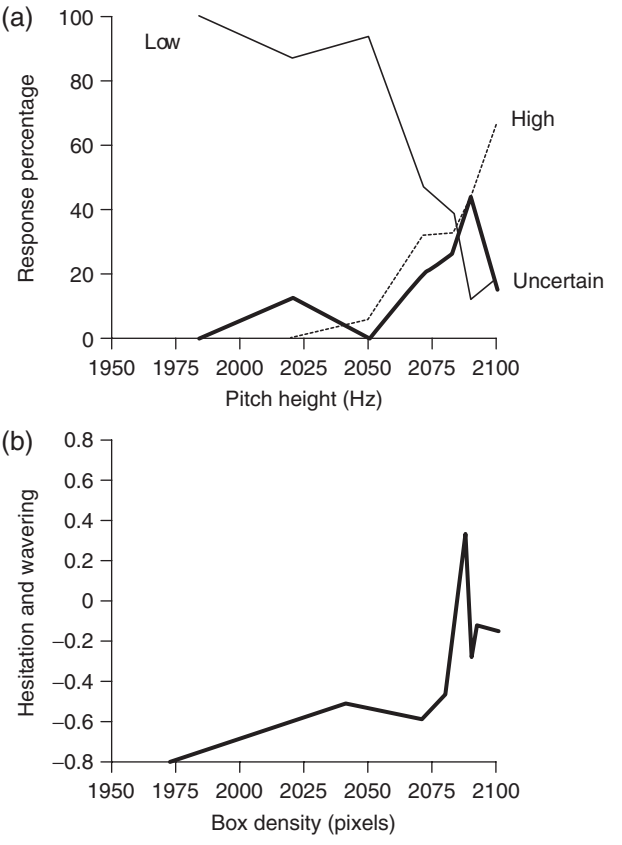


Fig. 1.1 a) Performance by a dolphin in the auditory discrimination of Smith et al. (1995). The horizontal axis indicates the frequency (Hz) of the trial. The 'high' response was correct for tones at 2100 Hz—these trials are represented by the rightmost data point for each curve. All lower-pitched tones deserved the 'low' response. The solid line represents the percentage of trials receiving the 'uncertain' response at each trial level. The percentages of trials ending with the 'high' response (dashed line) or 'low' response (dotted line) are also shown. b) The dolphin's weighted overall ancillary uncertainty behaviours (hesitancy, slowing, wavering) for tones of different frequencies (Hz). Reproduced from 'The Uncertain Response in the Bottlenosed Dolphin (*Tursiops truncatus*)', by J.D. Smith, J. Schull, J. Strote, K. McGee, R. Egnor, and L. Erb, 1995, *Journal of Experimental Psychology: General*, 124, pp. 391, 402 © 1995, The American Psychological Association with permission.

(Staddon et al. 2007, 2009; Carruthers 2008; Smith et al. 2008; Carruthers and Ritchie Chapter 5, this volume; Crystal Chapter 2, this volume), and these must be addressed for the field to move forward.

The first issue concerns the associative weight that the uncertainty response might take on after sufficient training. In Smith et al. (1997), as well as in many subsequent studies (e.g. Inman and Shettleworth 1999; Hampton 2001; Foote and Crystal 2007; Kornell et al. 2007; Suda-King 2008; Sutton and Shettleworth 2008; Fujita 2009; Roberts et al. 2009), the uncertainty response was rewarded. Although this reward was often less desirable than rewards for the primary responses, it still granted associative strength and attractiveness to a response that was supposed to only

indicate uncertainty. Although this possibility does not entirely discount these studies, modelling analysis suggests that animals might use the uncertainty response in some studies based purely on learned associations and not on uncertainty monitoring (see Smith et al. 2008).

A second issue is that the stimuli in Smith et al. (1995, 1997) were concrete and fixed. Animals experienced these auditory or visual stimuli many times, and might have learned that certain stimuli were more often followed by penalty timeouts than were other stimuli. For these stimuli, even an unrewarded uncertainty response presented a wholly superior outcome compared to the penalty that had come to be expected. That is, when used only on the most difficult trials, the response avoided a penalty, which is obviously superior to receiving a penalty. In some sense this could be considered a 'reward', because it produces a favourable outcome when used on trials in which the animal would otherwise receive a penalty. However, it differs in the traditional sense of reward because, when pressed, the traditional food/penalty outcomes are simply avoided. Furthermore, getting this 'reward' assumes that the animal can assess difficulty (based on its subjective impression of the task), which may suggest it has some access to mental states. Still, because the uncertainty response produced this more favourable outcome its use might be a reaction dictated by the stimulus properties rather than an introspective judgement of uncertainty.

These two problems were exacerbated by a third issue, which was that animals have traditionally been given trial-by-trial feedback in uncertainty-monitoring paradigms. This transparent feedback might have allowed animals to immediately determine whether their response was correct. Animals might come to learn that it is beneficial to avoid certain stimuli or certain situations by tracking their reinforcement history when responding to those stimuli or situations. Using the uncertainty response when presented with the least-often-rewarded stimuli would be very efficient. In fact, comparative researchers rely on the assumption that animals are always trying to maximize their food rewards, and receiving transparent trial-by-trial feedback gives them the ability to do so without relying on mental states or uncertainty. In contrast, deferred and rearranged feedback (see 'Opaque reinforcement' section) decreases, and possibly prevents, the associative system from learning the relationship between stimuli, responses, and outcomes. Thus, use of the uncertainty response under this paradigm cannot easily be explained by simply associative processes. However, the initial studies reported discussed here had only used transparent feedback.

All three of these problems were explored mathematically by Smith et al. (2008). Smith and colleagues created response profiles for three possible strategies that an animal might use in the kind of task described earlier: 1) a metacognitive strategy, where secondary opt-out responses were motivated by simulated subjective uncertainty; 2) an associative strategy where the secondary opt-out response was rewarded and trial-by-trial feedback was given; and 3) a stimulus-avoidance strategy where the simulated animal tracked reinforcement history and used the secondary opt-out response to avoid stimuli that it had been punished for previously. It was found that all three strategies produced nearly identical simulated behaviour in the task. That is, given trackable reinforcement and concrete stimuli, it was impossible to tell whether an animal was responding uncertain based on a subjective evaluation of its mental states or based on learned associative cues.

This analysis called into question many comparative claims about the metacognitive abilities of animals. Similar modelling from signal detection and behavioural economic perspectives (e.g. Staddon et al. 2007, 2009; Crystal and Foote 2009; Jozefowiez et al. 2009) have also found that the seemingly metacognitive results might be accounted for by lower-level associative processes. However, it is important to keep in mind that these models attempt to mathematically describe the data in the simplest way possible, using parameters that are tied to data points rather than mental processes. If an experimental task denied the animal access to clear non-mental, objective

cues or signals such as those that underlie associative learning, but the animals still preserved their uncertainty responding behaviour, then the model's low-level interpretations would not be applicable. Accordingly, a second generation of comparative enquiries into animal metacognition has attempted to create tasks that might better tap animals' potential capacity for uncertainty monitoring while being outside the bounds of traditional associative explanations.

New paradigms: abstract, opaquely reinforced, and naturalistic

Abstract judgements

In an early demonstration of uncertainty monitoring in abstract situations, Shields et al. (1997) showed that rhesus monkeys adaptively used the uncertainty response in a same–different task. Monkeys were asked to decide whether two boxes were equally pixilated or not. The actual level of pixilation (the stimulus values) of the boxes did not matter at all; all that mattered was whether the two boxes were the same or different. In this way, it was impossible to learn to associate certain stimulus values with penalties or rewards because the actual stimuli were not the determining factor in the primary discrimination. Monkeys still declined trials that featured the most difficult low-disparity pairs of boxes. Similar results have been found in the abstract domain of numerosity judgements (Beran et al. 2006). Even when the relationship between stimuli or an abstract feature of the stimuli such as number is the crucial factor—not the stimuli themselves—and there are no cues that can be associatively conditioned towards, animals still make adaptive uncertainty responses.

Another way in which researchers move beyond stimulus-based cues is by using tasks that measure metamemory. In these tasks, the reward contingencies are not tied to the to-be-remembered items themselves, but rather to the strength of the participant's memory. Stimuli may be different for each trial and may never repeat, thus preventing associations or avoidance biases. Smith et al. (1998) gave rhesus monkeys a serial-position metamemory task. Monkeys saw a series of images and were then shown a probe item and asked if it had been one of the items in the series. Not surprisingly, animals were best at making these judgements when the item was near the beginning or end of the series. They also declined trials featuring probes from the middle, and therefore hardest, positions, suggesting that they understood that the strength of their memory for these items was weakest. Hampton (2001) similarly found that monkeys would decline trials after longer delays in a matching-to-sample task compared to shorter delays, and Kornell et al. (2007) obtained similar results in a token-economy metamemory task. All of these paradigms moved animals away from stimulus-contingency associations by making the primary judgements be about memory rather than the properties of currently present stimuli.

In an interesting use of the metamemory paradigm, Washburn et al. (2009) used transcranial magnetic stimulation (TMS) to selectively interfere with some memories in a matching-to-sample task. A monkey was briefly shown a shape in the periphery of his left or right visual field. This created a visual image in the contralateral cerebral hemisphere that could be retained through an interval. TMS was applied on some trials to the same hemisphere as the visual image, or to the opposite hemisphere. Fig. 1.2 shows that on trials without TMS or with TMS applied to the opposite hemisphere from the visual stimulation, primary responding was accurate and uncertainty responding was low. But when TMS was applied to the same hemisphere as the visual stimulation, uncertainty responding was significantly higher. This suggests that the animal was able to discern when TMS had disrupted his visual memory for the event, and could respond accordingly. He knew when he remembered, and when TMS erased his memory he knew that he did not remember. Of course, we do not know how the monkey actually experienced these memory failures.

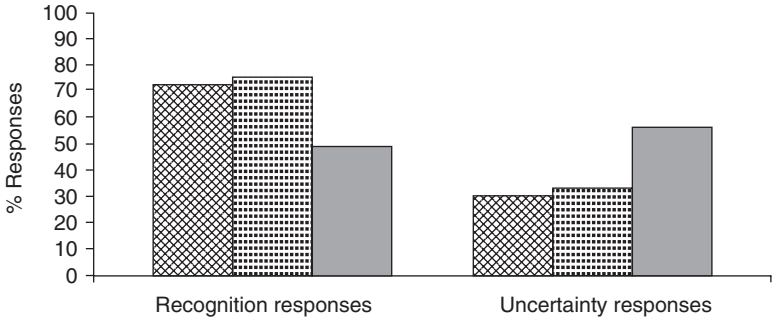


Fig. 1.2 Percentage of correct recognition responses and uncertainty responses for each condition in Washburn et al. (2009). Boxes with crosses indicate no TMS. Boxes with squares indicate TMS to the hemisphere opposite initial stimulus processing. Shaded boxes indicate TMS to the same hemisphere as initial stimulus processing. Reproduced from 'With his memory magnetically erased, a monkey knows he is uncertain' by D.A. Washburn, J.P. Gulledge, M.J. Beran, and J.D. Smith, 2010, *Biology Letters*, 6, pp. 160–2. © 2009 Royal Society Publishing with permission.

They may have been similar to the human experience of forgetting, they may have been phenomenologically profound, or they may have been entirely different. The study suggests only that the animal was monitoring his memory in some way.

Opaque reinforcement

In addition to making tasks more abstract, researchers have also sought to dissociate performance from reward contingencies. Washburn et al. (2006) found that monkeys given Harlow's (1949) learning-set paradigm would respond uncertain to the first trial of each new task in order to get information about which choice was correct for the series. The behaviour transferred to many new stimulus sets despite the monkeys having no reinforcement history (or any training) on them. The monkeys learned that new stimuli presented particular difficulty on trial 1 because they could not know which stimulus was the S+ and thus were at risk for making errors.

Smith et al. (2006) asked monkeys to make responses to a sparse–dense discrimination under conditions of deferred and rearranged feedback. After completing four trials, monkeys would receive summary feedback—all rewards first, followed by all penalties. This situation is similar to some human exam environments, where summary scores are delivered only after the test is completed; judgements about the difficulty of each question must be made without the aid of reinforcement history for similar questions. Humans and monkeys had some idea about the material (i.e. what generally constituted as sparse and dense), but had no way of knowing how well they were doing at the task or where exactly the line between sparse and dense should be drawn. Smith et al. found that one monkey used the uncertainty response in this environment. An analysis of his results suggested that he was not tracking his reinforcement—for example, he chose to make primary responses to some stimulus levels that he often answered incorrectly—but he was consistently following his cognitive construal of the task. There was a strong relationship between the animal's decisional breakpoint and the proportion of uncertainty responses made.

This work was extended by Couchman et al. (2010) using a paradigm that more closely mirrored the structure of Smith et al. (1995, 1997) and also incorporated novel transfer tasks. Three rhesus monkeys were trained to respond under deferred and rearranged feedback, and were then

transferred to qualitatively new and different transfer tasks for which they had no reinforcement history or training. These included judgements of line lengths, continuity, ellipticity, angle, and others. Monkeys (and humans, too) were given feedback only for the most extreme stimulus values—e.g. the shortest and longest line—and were then brought to their perceptual threshold under deferred feedback. Because reinforcement was opaque, they had no way of associating primary response contingencies to specific stimulus values. At the same time, associative cues from previous tasks were unhelpful, because each new stimulus continuum had a different perceptual threshold for each participant. Humans and three monkeys adaptively declined the most difficult trials in several new tasks. Fig. 1.3 shows the performance of one monkey in three novel transfer tasks. Notice that for each task, uncertainty responding peaks close to the point at which the primary responses converge, suggesting that the animal was adaptively declining the most difficult trials. It is an important fact that this response pattern was different for each new task, eliminating the possibility that uncertainty responding was a carryover effect or the result of previous learning. This suggests that the humans' and monkeys' uncertainty-monitoring ability is not dependent on any particular stimulus continua or paradigm, but instead is probably a free-floating ability that evolved to respond to general uncertainty.

Naturalistic paradigms

Researchers have also used naturalistic paradigms, particularly with chimpanzees and orangutans, because they require less training and thus may not be subject to associative criticisms. In Call and Carpenter (2001), human children, chimpanzees, and orangutans were asked to choose a tube with a possible food reward inside (see also Call Chapter 4, this volume). The participants either saw or did not see the food reward being hidden. When they saw the food reward being hidden, they immediately chose the correct tube. When the food was hidden out of sight, they sought additional information by looking into each tube before making a choice. This suggests that they knew when they had and had not seen the food being hidden. Hampton et al. (2004) found similar results in rhesus monkeys. Suda-King (2008) found that orangutans would immediately choose a tube that they had seen baited with two grapes, but would choose to take a one-grape guarantee rather than risk choosing a tube they had not seen being baited. In both cases, the uncertainty monitoring behaviour emerged with very little training and was tied to memory experiences in a naturalistic setting.

These paradigms open new avenues into comparative metacognition and raise several important questions. Clearly the evidence presented so far indicates that animals can use an escape response or hint-seeking behaviour when they face difficulty or lack of information. However, the overwhelming success of a dolphin, rhesus monkeys, and apes in these paradigms might itself be concerning. Perhaps some yet-undiscovered associative or behavioural trick allows animals to make uncertainty responses without actually having access to their mental states. Although the new paradigms described earlier have gone a long way to rule out the effects of associative cues, carryover from previously learning, and stimulus aversion, there is always the possibility that another simple behavioural strategy is lurking in the darkness. To explore this possibility, it is important to highlight some influential contrast cases in which associatively-capable species fail to use the uncertainty response.

What we learn from failures

If the use of the uncertainty response was wholly dictated by any sort of associative strategy, then it stands to reason that all creatures capable of building associations ought to also be capable of using the uncertainty response. This is true regardless of whether their behaviour is motivated by

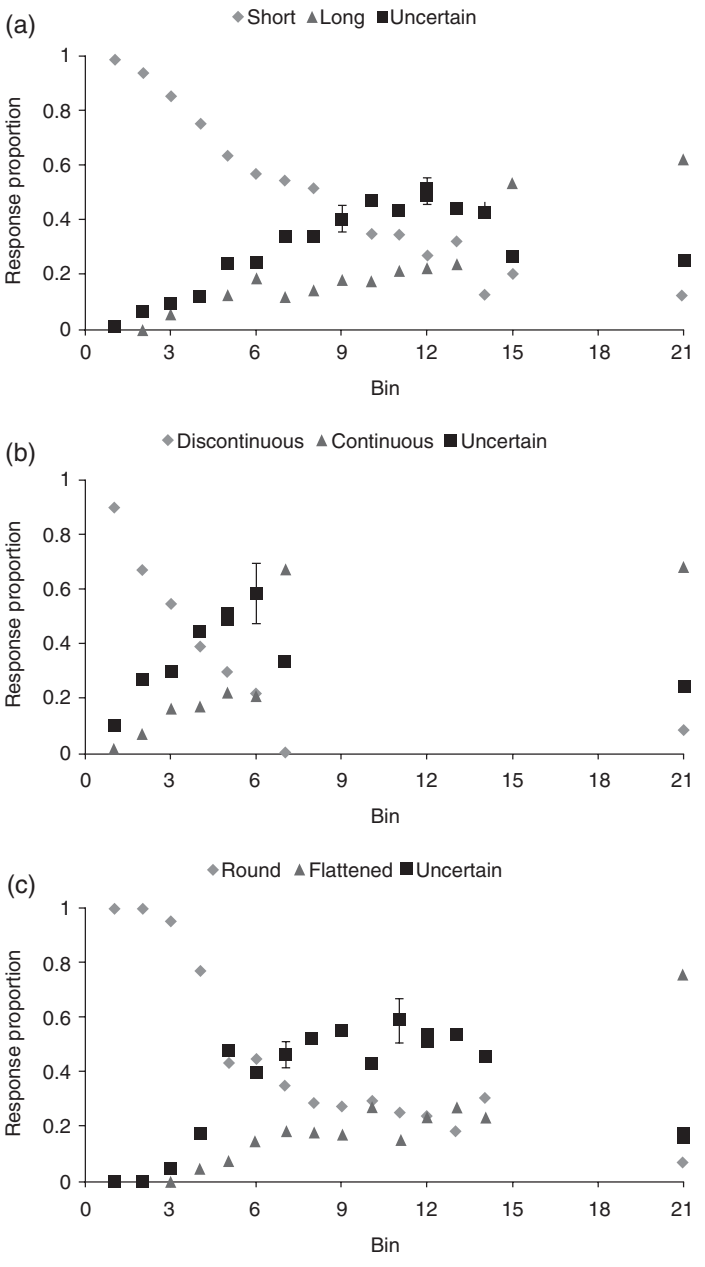


Fig. 1.3 Performance by one monkey in three novel visual discrimination tasks in Couchman et al. (2010). The horizontal axis indicates the stimulus levels. The long (a), continuous (b), or flattened (c) response was correct for bin 21—these trials are represented by the rightmost data point. All lower bins deserved the short (a), discontinuous (b), or round (c) response. Grey diamonds and triangles indicate these primary responses. Black squares indicate uncertainty responses. Error bars show 95% confidence intervals for selected bins. Reproduced from 'Beyond stimulus cues and reinforcement signals: A new approach to animal metacognition' by J.J. Couchman, M.V.C. Coutinho, M.J. Beran, and J.D. Smith, 2010, *Journal of Comparative Psychology*, 124, pp. 356–68 © 2010 The American Psychological Association with permission.

stimulus-aversion strategies, behavioural economic strategies, Carruthers' (2008) desire-strength system, or any other sort of first-order system. However, on the contrary, comparative cognition has uncovered several species that are perfectly capable of first-order associative processing—e.g. pigeons and capuchin monkeys—but are nonetheless incapable of making adaptive uncertainty responses. The failures of these otherwise very capable species serve as a sharp contrast to the success of the dolphin, rhesus monkeys, and apes.

Inman and Shettleworth (1999) and Sutton and Shettleworth (2008) found that when difficulty increased in a metamemory experiment, pigeons did not increase their uncertainty responding. This was true even when the uncertainty response brought a small reward. Pigeons also failed to use a hint-seeking response that would have provided information necessary to complete a matching-to-sample task (Roberts et al. 2009). Although pigeons are certainly capable of tracking their feedback, learning to avoid certain stimuli, and the like, they were still unable to use the uncertainty response (but see Fujita et al. Chapter 3, this volume).

Capuchin monkeys have also failed to show uncertainty-monitoring capabilities. In tasks similar to Call and Carpenter (2001), capuchins largely failed to visually search tubes that they had not seen being baited with food (Basile et al. 2009). They chose without regard to what they had seen or not seen. Capuchins also perform odd or unnecessary search behaviours such as visually inspecting transparent tubes (Paukner et al. 2006).

Beran et al. (2009) gave capuchins a sparse–dense uncertainty-monitoring task that had been used with rhesus monkeys and humans. Capuchins performed similarly to humans and rhesus monkeys on the primary discriminations, but essentially did not use the uncertainty response at all. Even when the penalty for an incorrect answer was more than quadrupled, capuchins still did not use the uncertainty response to avoid answering difficult trials. In a separate condition, capuchins were given the option of using a 'middle' response that was rewarded for the section of the continuum that was associated with uncertainty-response use in humans and rhesus monkeys. This 'middle' response, a primary response for the task, was easily mastered by the capuchins. Yet despite the fact that using the uncertainty response on the same section of the continuum would have been highly beneficial, capuchins in the uncertainty condition were generally unable or unwilling to use it (see Smith et al. 2009b).

The failure of pigeons and capuchins underscores the psychological status of the uncertainty response. Any theory suggesting that uncertainty responding is motivated by non-metacognitive processes would need to explain why pigeons and capuchins lack the ability. Obviously this is no easy task. In a variety of tasks, pigeons and capuchins have been shown to be sensitive to low-level conditioning processes, response-strengths, reinforcement histories, and reward maximization. They are not at all lacking in these abilities, but do seem to lack the higher-level ability to monitor their own uncertainty in a way that exceeds first-order processing. Of course there are many models of learning that might explain the animals' behaviour. However, none to date has been able to explain both the capuchin and pigeon failures and the rhesus monkeys' success without using some form of uncertainty monitoring.

Implications for metacognition

It is difficult to deny that animals make escape responses or information-seeking responses in a variety of paradigms and that they do so without relying entirely on reinforcement cues or associative strategies. No mathematical model we are aware of can explain the reported performances in deferred and rearranged feedback environments, and certainly none can explain why pigeons and capuchins fail to respond uncertain even when every necessary associative cue is present. The animal evidence accumulated thus far has led many to believe that at least some animals do have

some capacity for metacognition (Metcalfe 2008; Sutton and Shettleworth 2008; Fujita 2009; Roberts et al. 2009; Smith 2009).

Very few alternative possibilities remain. It is possible that animals might be conditioned to background aspects of the tasks, and might simply be responding to cues that researchers have not yet discovered. However, it is difficult to accept this possibility given that a dolphin, rhesus monkeys, and apes have shown evidence for metacognition in several different research laboratories that employed several very different paradigms. It is difficult to imagine a single cue (or set of cues) that can fully explain animals' performance and were common to all of these settings. Also, it is difficult to imagine why pigeons and capuchins that would have equal access to these background cues and are perfectly capable of associative learning did not use them to make uncertainty responses in any paradigm.

The experiments described previously in this chapter have eliminated every such cue that has been suggested, including associative learning, stimulus aversion, perceptual information from specific stimuli, reinforcement tracking, experience with the task (via first-trial analyses), all aspects of computerized testing (via naturalistic experiments), and all aspects of naturalistic testing (via computerized experiments). Although it is true that each particular experiment might be subject to some associative loophole, one can easily find experiments where that loophole is closed. No experiment is perfect, but taken together the body of evidence favours metacognitive explanations. Although scientific intuition might suggest that associative explanations are more parsimonious than metacognitive explanations, one must keep in mind that metacognition explains all of the results with one relatively simple process—one that we know humans have, and one that evolution probably selected for—while alternative explanations require different strategies to explain each result.

Furthermore, although mathematical models (Staddon et al. 2007, 2009; Smith et al. 2008) and logic systems (Carruthers 2008) have raised some concerns over the interpretation of the uncertainty response, it is important to note that no experiment has ever confirmed an alternative explanation for dolphins, monkeys, or apes while disconfirming the metacognitive explanation. And, some remaining alternative explanations have not yet been tested. Though this is the general process of science, we believe it is important to note that, to date, actual experimental results for these species have alleviated many concerns about non-metacognitive explanations. If any associative model, or for that matter any alternative explanation were valid, then it ought to be a simple matter to give animals one task in which the proposed behavioural cue was available, and another in which it was not. Animals would adaptively use the uncertainty response when the cue was present, and would fail to respond uncertain when it was not. Many such experiments are described in earlier sections, and in all cases the rhesus monkeys used the uncertainty response appropriately regardless of whether or not the proposed cue was available. We always welcome new proposals, but have not found any that could explain all of the results described previously; the models focus on one or two paradigms each, but their approaches do not seem to generalize to all animal paradigms.

Does this mean that some animals have human-like conscious metacognition? In the tasks described earlier, humans making uncertainty responses or hint-seeking behaviours would certainly claim to be motivated by a conscious awareness of their uncertainty. It may be that some animals are on the verge of similar awareness (Smith 2009; Smith et al. 2009a). Or, their metacognition may lack human-like awareness (Carruthers 2008). Awareness is an elusive issue, but it is important for future work in comparative metacognition to explore the possibility of explicit cognition, explicit metacognition, and other executive processes that might provide clues towards determining the level of metacognitive awareness that some animals possess. This is a critical component for determining the nature and evolutionary history of the reflective mind (Humphrey 1976; Gallup et al. 1995) and understanding the development of human consciousness.

Finally, one of the most difficult questions concerns the representational nature of animal metacognition. Some animals know when they do not know. But, do they know 'I know' or 'I remember' or 'I believe' or some non-linguistic equivalent? Do they represent their knowledge as a belief or memory state, or do they have some more primitive way of monitoring their minds? It is important to consider that these animals have come to have metacognition while apparently lacking mindreading capabilities that are often associated with the representation of mental states (Carruthers 2009; Couchman et al. 2009; though see Clements and Perner 1994). How is this possible?

We know that some animals consistently and adaptively use the uncertainty response and seek additional information when their memory is insufficient. We know that, despite many attempts, no experiment has discovered an associative, economic, first-order, or any other explanation for the results that discounts the possibility of a metacognitive explanation at the same time. We know that these facts apply to animals that apparently lack mindreading abilities. Does this mean that metacognition is not dependent on mindreading, or perhaps that it evolutionarily preceded it? We believe it does, based on empirical evidence. We recognize that others have different views on the nature of mindreading and metacognition, but the burden of proof is on them to experimentally falsify the current evidence or provide counter-evidence. Failing that, any theory or explanation of human or animal metacognition must be in line with the evidence described in this chapter.

It has been suggested that, despite this evidence, animals' behaviour in this task might not 'count' as metacognition because it lacks some metarepresentational property (Carruthers 2008, 2009). There are a variety of reasons to think that animals' behaviour in uncertainty-monitoring tasks should be considered metacognitive. It has long been suggested that indeterminacy leads to controlled, even executive processing (Shiffrin and Schneider 1977). Stimuli at the perceptual threshold are known to map poorly onto behavioural responses and elicit more sophisticated cognitive behaviours. Even Carruthers (2008), when describing a supposedly non-metacognitive explanation, grants an extra, slower, more informed, and ultimately decisional gatekeeper mechanism to account for these situations. James (1890/1952), Dewey (1934/1980), Karoly (1993), and others all describe situations in which mental conflict results in heightened self-awareness and initiates higher-level processes of self-regulation. Tolman (1927) even considered hesitation/wavering behaviours (e.g. Fig. 1.1b) to be a way in which behaviourists could operationalize animal consciousness. Of course, many of these self-regulation behaviours in humans appear to have verbal correlates, and animals cannot give verbal reports to reinforce our findings. Still, it would be unusual to claim that in humans the experience of uncertainty-monitoring is intertwined with sophisticated cognitive processing, representation, and consciousness, while in animals that display isomorphic behaviours—and that clearly would have faced situations where such mental processes were evolutionarily beneficial—it is something else entirely.

It is important to also consider the implications that pigeon and capuchin failures have for the question of representation. Although there might be strong reasons to believe that uncertainty states are directly tied to higher-level processing, it might be that uncertainty responding is caused by the uncertainty state itself (not by awareness of being in that state). This explanation would require one to believe either that pigeons and capuchins do not experience uncertainty or that their experience of uncertainty does not lead to any beneficial behaviours. Both are exceedingly unlikely, given that significant uncertainty is common in nature (Griffin 2003). If the uncertainty state alone was sufficient for uncertainty responding, then virtually all animals should be capable of making the response in difficult situations. Because they do not, the most likely possibility is that something more is required. By contrast, the metacognitive explanation accounts for this difference. Capuchins and pigeons probably experience uncertainty, but lack the awareness

necessary to respond adaptively in uncertainty monitoring paradigms. This is not to say they lack any awareness of uncertainty, only that they lack the degree to which we have been able to test so far (but see Fujita et al. Chapter 3, this volume). Rhesus monkeys are aware of their uncertainty and can respond adaptively. Capuchins and pigeons are not (as) aware and cannot respond adaptively. This seems to be a better explanation than the 'state itself' explanation, which holds that these species all experience uncertainty, but for unknown reasons some cannot respond to it adaptively. It is possible that a critical non-metacognitive process differentiating these species exists but has not yet been discovered, but we believe that metacognitive differences are a better explanation than an unknown possibility.

It is of course vitally important to determine the actual content of animals' metacognitive states and determine exactly what kind of extra processing is required to facilitate uncertainty responding. Are some animals representing their beliefs and desires as mental states or doing some more implicit form of self-monitoring? In the developmental theory of mind literature, pre-linguistic subjects are thought to demonstrate nascent sensitivity to others' beliefs even if they lack explicit theory of mind (Low 2010). The same sort of nascent sensitivity to one's own beliefs might be the motivating factor behind animal metacognition, and as such this more primitive understanding might underlie metarepresentational metacognition in adult humans. Or, animals might be metarepresenting their beliefs and researchers have simply not yet discovered a way to demonstrate it. A great deal of further work is needed to fully investigate these issues, but the current state of our field shows some evidence for animal metaminds and suggests that the ability to understand one's own mental states is not unique to humans.

The question we face is not whether some animals have or lack metacognition, nor is it whether certain conditions are somehow required for a process to count as metacognitive. These all-or-none distinctions are almost exclusively semantic and do little to further our understanding of the reflective mind—they haggle over the price rather than examine the product. The great question is where animal metacognition falls on the spectrum ranging from primitive cognition to full human consciousness, and what it can tell us about the nature and gradual evolution of human awareness. In asking that question, we can begin to examine whether the nature of metacognition can be procedural and/or conceptual (Proust Chapter 14, this volume), whether it is different from metarepresentational abilities (Carruthers and Ritchie Chapter 5, this volume), or whether it has several different varieties (Perner Chapter 6, this volume). These important questions can be asked if we understand the nature of metacognition in different species, and in some cases can only be answered by examining human and animal performances in the paradigms described here and in the other chapters of this volume.

References

Basile, B.M., Hampton, R.R., Suomi, S.J., and Murray, E.A. (2009). An assessment of memory awareness in tufted capuchin monkeys (*Cebus apella*). *Animal Cognition*, 12, 169–80.

Benjamin, A.S., Bjork, R.A., and Schwartz, B.L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metacognitive index. *Journal of Experimental Psychology: General*, 127, 55–68.

Beran, M.J., Smith, J.D., Redford, J.S., and Washburn, D.A. (2006). Rhesus macaques (*Macaca mulatta*) monitor uncertainty during numerosity judgments. *Journal of Experimental Psychology: Animal Behavior Processes*, 32, 111–19.

Beran, M.J., Smith, J.D., Coutinho, M.V.C., Couchman, J.J., and Boomer, J. (2009). The psychological organization of 'uncertainty' responses and 'middle' responses: A dissociation in capuchin monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 371–81.

Call, J. and Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition*, 4, 207–20.

Carruthers, P. (2008). Meta-cognition in animals: a skeptical look. *Mind and Language*, 23, 58–89.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32, 121–38.

Clements, W.A. and Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, 9, 377–95.

Couchman, J.J., Coutinho, M.V.C., Beran, M.J., and Smith, J.D. (2009). Metacognitionis prior. *Behavioral and Brain Sciences*, 32, 142.

Couchman, J.J., Coutinho, M.V.C., Beran, M.J., and Smith, J.D. (2010). Beyond stimulus cues and reinforcement history: a new approach to animal metacognition. *Journal of Comparative Psychology*, 124, 356–68.

Crystal, J.D. and Foote, A.L. (2009). Metacognition in animals. *Comparative Cognition and Behavior Reviews*, 4, 1–16.

Dewey, J. (1934/1980). *Art as experience*. New York: Perigee Books.

Flavell, J.H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34, 906–11.

Foote, A. and Crystal, J. (2007). Metacognition in the rat. *Current Biology*, 17, 551–5.

Fujita, K. (2009). Metamemory in tufted capuchin monkeys (*Cebus apella*). *Animal Cognition*, 12, 575–85.

Gallup, G.G. (1982). Self-awareness and the emergence of mind in primates. *American Journal of Primatology*, 2, 237–48.

Gallup, G.G., Povinelli, D.J., and Suarez, S.D. (1995). Further reflections on self-recognition in primates. *Animal Behaviour*, 50, 1525–32.

Griffin, D.R. (2003). Significant uncertainty is common in nature. *Behavior and Brain Sciences*, 26, 346.

Hampton, R.R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 5359–62.

Hampton, R.R., Zivin, A., and Murray, E.A. (2004). Rhesus monkeys (*Macaca mulatta*) discriminate between knowing and not knowing and collect information as needed before acting. *Animal Cognition*, 7, 239–54.

Harlow, H.F. (1949). The formation of learning sets. *Psychological Review*, 56, 51–65.

Humphrey, N.K. (1976). The social function of intellect. In P.P. Bates and R.A. Hinde (Eds.) *Growing points in ethology*, pp. 303–17. Cambridge: Cambridge University Press.

Inman, A. and Shettleworth, S.J. (1999). Detecting metamemory in nonverbal subjects: A test with pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 25, 389–95.

James, W. (1890/1952). *The principles of psychology*. Vol. 53, Great Books of the WesternWorld. Chicago, IL: University of Chicago Press.

Jozefowiez, J., Staddon, J.E.R., and Cerutti, D.T. (2009). Metacognition in animals: How do we know that they know? *Comparative Cognition & Behavior Reviews*, 4, 29–39.

Karoly, P. (1993). Mechanisms of self-regulation: A systems view. *Annual Review of Psychology*, 44, 23–52.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609–39.

Koriat, A. (2007). Metacognition and consciousness. In P.D. Zelazo, M. Moscovitch, and E. Thompson (Eds.) *The Cambridge handbook of consciousness*, pp. 289–325. Cambridge: Cambridge University Press.

Kornell, N., Son, L., and Terrace, H. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18, 64–71.

Low, J. (2010). Preschoolers' implicit and explicit false-belief understanding: Relations with complex syntactical mastery. *Child Development*, 81, 597–615.

Metcalfe, J. (2008). Evolution of metacognition. In J. Dunlosky and R. Bjork (Eds.) *Handbook of Metamemory and Memory*, pp. 29–46. New York: Psychology Press.

Metcalfe, J. and Kober, H. (2005). Self-reflective consciousness and the projectable self. In H.S. Terrace and J. Metcalfe (Eds.) *The missing link in cognition: Origins of self-reflective consciousness*, pp. 57–83. New York: Oxford University Press.

Morgan, C.L. (1906). *An introduction to comparative psychology*. London: Walter Scott.

Nelson, T.O. (Ed.) (1992). *Metacognition: Core readings*. Toronto: Allyn and Bacon.

Nelson, T.O. (1996). Consciousness and metacognition. *American Psychologist*, 51, 102–16.

Nelson. T.O. and Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–41.

Paukner, A., Anderson, J.R., and Fujita, K. (2006). Redundant food searches by capuchin monkeys (*Cebus apella*): A failure of metacognition? *Animal Cognition*, 9, 110–17.

Roberts, W.A., Feeney, M.C., McMillan, N., MacPherson, K., Musolino, E., and Petter, M. (2009). Do pigeons (*Columba livia*) study for a test? *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 129–42.

Schwartz, B.L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin and Review*, 1, 357–75.

Shields, W.E., Smith, J.D., and Washburn, D.A. (1997). Uncertain responses by humans and rhesus monkeys (*Macaca mulatta*) in a psychophysical same-different task. *Journal of Experimental Psychology: General*, 126, 147–64.

Shiffrin, R.M. and Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84, 127–90.

Smith, J.D. (2009). The study of animal metacognition. *Trends in Cognitive Sciences*, 13, 389–96.

Smith, J.D., Schull, J., Strote, J., McGee, K., Egnor, R., and Erb, L. (1995). The uncertain response in the bottlenosed dolphin (*Tursiops truncatus*). *Journal of Experimental Psychology: General*, 124, 391–408.

Smith, J.D., Shields, W.E., Schull, J., and Washburn, D.A. (1997). The uncertain response in humans and animals. *Cognition*, 62, 75–97.

Smith, J.D., Shields, W.E., Allendoerfer, K.R., and Washburn, W.A. (1998). Memory monitoring by animals and humans. *Journal of Experimental Psychology: General*, 127, 227–50.

Smith, J.D., Shields, W.E., and Washburn, D.A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26, 317–73.

Smith, J.D., Beran, M.J., Redford, J.S., and Washburn, D.A. (2006). Dissociating uncertainty states and reinforcement signals in the comparative study of metacognition. *Journal of Experimental Psychology: General*, 135, 282–97.

Smith J.D., Beran, M.J., Coutinho, M.V.C., and Couchman, J.C. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin and Review*, 15, 679–91.

Smith, J.D., Beran, M.J., Couchman, J.J., Coutinho, M.V.C., and Boomer, J.B. (2009a). Animal metacognition: Problems and prospects. *Comparative Cognition & Behavior Reviews*, 4, 33–46.

Smith, J.D., Beran, M.J., Couchman, J.J., Coutinho, M.V.C., and Boomer, J.B. (2009b).The curious incident of the capuchins. *Comparative Cognition & Behavior Reviews*, 4, 47–50.

Staddon, J.E.R., Jozefowiez, J., and Cerutti, D. (2007). Metacognition: A problem not a process. *PsyCrit*, April 13, 1–5.

Staddon, J.E.R., Jozefowiez, J., and Cerutti, D. (2009). Metacognition in animals: How do we know that they know? *Comparative Cognition and Behavior Reviews*, 4, 29–39.

Suda-King, C. (2008). Do orangutans (*Pongo pygmaeus*) know when they do not remember? *Animal Cognition*, 7, 239–46.

Sutton, J.E. and Shettleworth, S.J. (2008). Memory without awareness: Pigeons do not show metamemory in delayed matching-to-sample. *Journal of Experimental Psychology: Animal Behavior Processes*, 34, 266–82.

Terrace, H.S. and Metcalfe, J. (Eds.) (2005). *The missing link in cognition: Origins of self-reflective consciousness*. New York: Oxford University Press.

Tolman, E.C. (1927). A behaviorist's definition of consciousness. *Psychological Review*, 34, 433–9.

Washburn, D.A., Smith, J.D., and Shields, W.E. (2006). Rhesus monkeys (*Macaca mulatta*) immediately generalize the uncertain response. *Journal of Experimental Psychology: Animal Behavior Processes*, 32, 85–9.

Washburn, D.A., Gulledge, J.P., Beran, M.J., and Smith, J.D. (2009). With his memory magnetically erased, a monkey knows he is uncertain. *Biology Letters*, 6, 160–2.

Chapter 2

Validating animal models of metacognition

Jonathon D. Crystal

Evolution of mind

The comparative analysis of metacognition is a pathway towards uncovering fundamental information about the evolution of mind. People can reflect on their own cognitive processes, an ability referred to as metacognition. People can assess their own knowledge. For example, I know that I am familiar with the geography of Canada. Hence, if asked about certain types of information (e.g. names of provinces, provincial capitals, major cities, etc.), I am confident about the answers that I may generate. Moreover, people can also assess their lack of knowledge. For example, I also know that I am not familiar with the geography of Belarus; if asked about basic information about that country, I am confident that I would not be able to generate any answers beyond guessing.

A substantial body of research has been directed towards the comparative analysis of metacognition, with a pace that accelerated as evidence emerged to suggest that some animals may possess metacognition. Although it is well established that people have metacognition (Dunlosky and Bjork 2008), studies of human metacognition can exploit both behavioural and subjective sources of information, and both sources have provided rich opportunities for generating hypotheses about human metacognition. However, the study of metacognition in animals focuses exclusively on behavioural sources of evidence because subjective sources cannot be evaluated in non-verbal animals. The goal of this chapter is to outline some ideas about what type of evidence is required to validate an animal model of metacognition. From the outset, the proposal that animals possess metacognition has been judged against the backdrop that basic forms of learning are a class of alternative explanations for putative metacognition data. For example, according to a *stimulus-response* hypothesis, an animal may learn to do a particular response in the presence of a specific stimulus. The next sections provide a brief review of examples of metacognition and delineate two hypotheses. The subsequent section outlines some examples of conflicting views about interpreting metacognition experiments. The final section outlines an approach towards resolving the conflict. Progress in the comparative analysis of metacognition is threatened by conflicting views about the standards required to document metacognition in animals.

Metarepresentations

To evaluate comparative metacognition, it is important to distinguish between representations and metarepresentations. Accordingly, the presentation of a stimulus gives rise to an internal representation of that stimulus (which will be referred to as a *primary* representation). Behaviours are frequently based on primary representations. For example, when presented with an item on a memory test, it is possible to evaluate familiarity with the item to render a judgement that the item is new or old. Metacognition involves a *secondary* representation which operates on a primary

representation (i.e. a metarepresentation). For example, a person can report on their knowledge that they do not know the answer to a question, and this awareness can impact behaviour (e.g. an appropriate action can be taken such as deferring until additional information is available). To validate an animal model of metacognition, we need a method that implicates the use of a secondary representation. Otherwise, how can we be certain that performance is not based on a primary representation?

Carruthers (2008) distinguished between first-order explanations and metacognition. According to Carruthers, first-order explanations are 'world-directed' rather than 'self-directed.' First-order explanations are representations about stimuli in the world (e.g. a belief about the world), whereas metacognition involves representations about beliefs (e.g. knowing that you hold a particular belief) according to this view. Carruthers argued that putative metacognitive phenomena in animals may be explained in first-order terms.

Examples of putative metacognition data

Two examples of putative metacognition data from rhesus monkeys (*Macaca mulatta*) (Hampton 2001) and rats (*Rattus norvegicus*) (Foote and Crystal 2007) are described as follows. Hampton (2001) trained monkeys in a matching to sample procedure (i.e. reward was contingent on selecting the most recently seen image from a set of distracter images) using daily sets of four clip-art images. The procedure is outlined in Fig. 2.1. Foote and Crystal (2007) trained rats to categorize noise durations as short or long (i.e. reward was contingent on judging the four shortest and four longest durations as short and long, respectively) using a set of eight durations. The procedure is outlined in Fig. 2.2. The two experiments have a number of common features. Before taking some tests, the animals were given the opportunity to decline the test. On other trials, the animals were not given the option to decline it. Accurate performance on the test yielded a valuable reward, whereas inaccurate performance resulted in no reward. Declining a test yielded a less valuable (but guaranteed) reward. The decline rate increased as a function of difficulty (longer retention intervals for monkeys or proximity to the subjective middle of short and long durations for rats) and accuracy was lowest on difficult tests that could not be declined.

The data in Figs 2.1 and 2.2 from a monkey and rats are similar. There are two important features of the data. First, difficult tests were declined more frequently than easy tests, which may suggest that the animals adaptively used the decline response. Second, the decline in accuracy as a function of difficulty was more pronounced with tests that could not be declined (forced test) compared to tests that could have been declined (choice tests); the latter pattern of data is referred to as the chosen–forced performance advantage, which appears to emerge as a function of task difficulty.

Two hypotheses: basic learning mechanisms and metacognition

Two types of hypotheses may be offered to explain the data shown in Figs 2.1 and 2.2. According to a metacognition perspective, animals have metarepresentations. An animal with metacognition should have higher accuracy when it chooses to take a test relative to its accuracy when forced to take the test. The rationale for this hypothesis follows: if the animal 'knows that it does not know' the correct response, then it will decline the test; moreover, being forced to take a test is likely to degrade performance because forced tests include trials that would have been declined had that option been available. Inman and Shettleworth (1999) introduced the idea that it is critical to assess accuracy with and without the opportunity to decline difficult tests and first

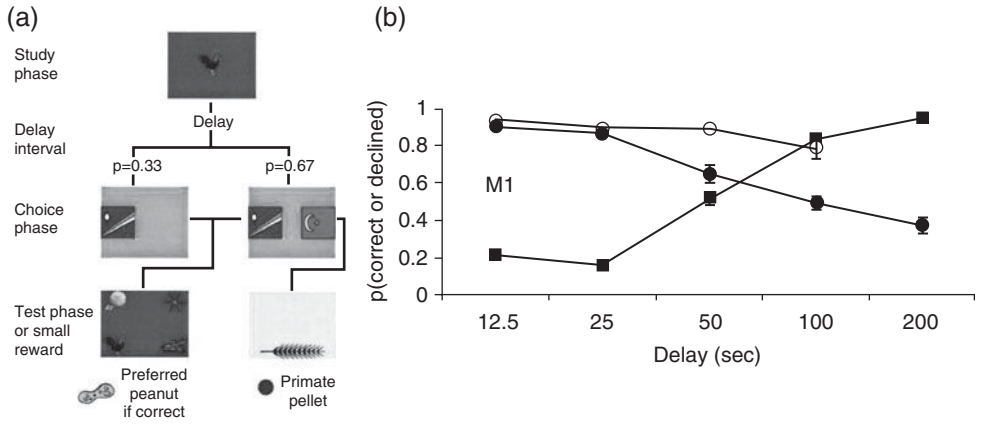


Fig. 2.1 Schematic representation of design of study and data. a) Procedure for monkeys (Hampton 2001): after presentation of an image to study and a brief retention interval, a choice phase provided a choice to take or decline a memory test; declining the test produced a guaranteed (but less preferred) reward than was earned if the test was selected and answered correctly (test phase); when a distracter image was selected in the memory test, no food was delivered. Items were selected by contacting a touchscreen. b) Data (Hampton, 2001): performance from a monkey that both used the decline response to avoid difficult problems (i.e. after a relatively long retention interval) and had a chosen–forced performance advantage (i.e. accuracy was higher on trials in which the monkey chose to take the test compared with forced tests, particularly for difficult tests). Filled squares represent the proportion of trials declined, and filled and unfilled circles represent proportion correct on forced and chosen trials, respectively. Error bars represent standard errors. Adapted from Hampton, R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 5359–62, figures 1 and 4. © 2001 National Academy of Sciences, U.S.A., with permission.

proposed that an animal without metacognition would have the same level of accuracy in forced test and choice test conditions.

Next an alternative to metacognition is outlined. According to a basic learning perspective, fundamental principles of learning may explain putative metacognition data from animals. According to this view, principles known to exist in animals (e.g. memory, generalization, resolution of response competition, habit formation, etc.) may be evaluated to determine if these principles are sufficient to explain data such as those shown in Figs 2.1 and 2.2. A theoretical model and simulations may be used to evaluate suitability to explain data (Church 1997). A theoretical model specifies processes that are proposed to explain how, for example, an animal generates behaviour when confronted with a procedure. The model might specify quantitative, psychological, or biological levels of implementation, and it can be explored in simulations by adding variability to specified parts of the model. Simulations can be compared to data to determine if the model provides a reasonable description of data. Importantly, from the basic learning perspective, metarepresentations are not proposed, thereby allowing an assessment of how much data can be explained by basic mechanisms of learning. The basic learning perspective may be considered a low-level alternative to metacognition in the sense that the basic learning perspective uses primary representations without application of secondary representations. Navigating the interpretation of data is further complicated by the possibility that a combination of basic learning mechanisms and metacognition may be involved.

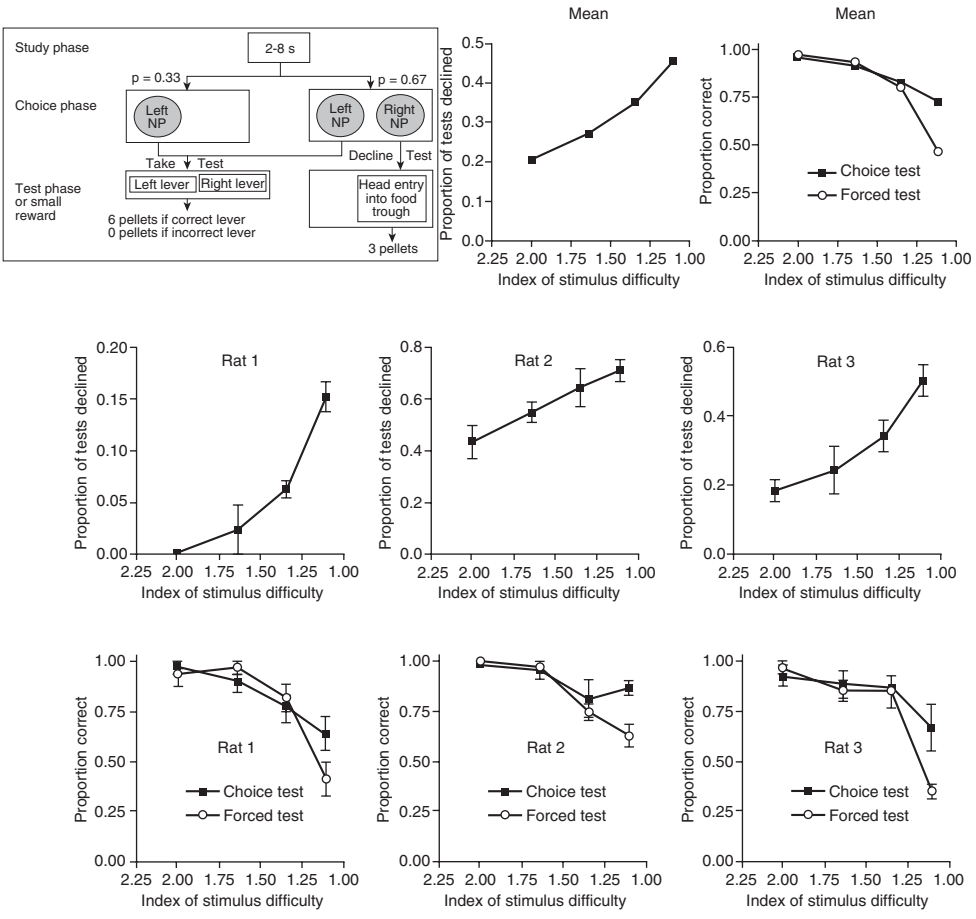


Fig. 2.2 Procedure for rats (top left panel; Foote and Crystal 2007): after presentation of a brief noise (2–8 s; study phase), a choice phase provided a choice to take or decline a duration test; declining the test produced a guaranteed (but smaller) reward than was earned if the test was selected and answered correctly (test phase). The shading indicates an illuminated nose-poke (NP) aperture, which was used to decline or accept the test. Data (Foote and Crystal, 2007): performance from three rats (bottom panels) and the mean across rats (top-middle and top-right panels). Difficult tests were declined more frequently than easy tests; difficulty was defined by proximity of the stimulus duration to the subjective middle of the shortest and longest durations. The decline in accuracy as a function of stimulus difficulty was more pronounced when tests could not be declined (forced test) compared to tests that could have been declined (choice test). Error bars represent standard errors. Adapted from *Current Biology*, 17(6), Allison L. Foote and Jonathon D. Crystal, Metacognition in the Rat, pp. 551–5, Copyright (2011), with permission from Elsevier.

Recent quantitative modelling by Smith and colleagues (2008) shows that a basic learning perspective can produce both apparently functional use of the decline response and the chosen–forced performance advantage without hypothesizing metacognition. Consequently, the modelling suggests that avoiding difficult problems and the chosen–forced performance advantage may not be sufficient to document metacognition.

Smith and colleagues (2008) used basic learning principles in their model. They proposed that rewarding the decline response produces a low-frequency tendency to select that response independent of the stimulus in the primary discrimination. Importantly, Smith et al. proposed that the decline response has a constant attractiveness across the stimulus continuum; constant attractiveness means that the tendency to produce the response is constant across stimulus conditions. We refer to this class of threshold explanations as a *stimulus-independent* hypothesis to contrast it with a stimulus-response hypothesis (i.e. 'in the presence of a particular stimulus, do a specific response'). For the primary discrimination, Smith et al. assumed exponential generalization decrements for an anchor stimulus in a trained discrimination. Such exponential decay functions have extensive empirical and theoretical support (Shepard 1961, 1987; White 2002); exponential decay is also commonly used to model a fading memory trace (Shepard 1961; Sikström 1999; Anderson 2001; Killeen 2001; Sargisson and White 2001, 2003, 2007; White 2001, 2002; Wixted 2004). According to this proposal, the primary discrimination and the decline option give rise to competing response-strength tendencies, and the behavioural response on a given trial is the one with the highest response strength (i.e. a winner-take-all response rule). A schematic of the model appears in Fig. 2.3a. Simulations document that the model can produce both apparently adaptive use of the decline response to effectively avoid difficult problems and a chosen–forced performance advantage that emerges as a function of task difficulty (Fig. 2.3b). Note that both putative metacognition data patterns are produced by the simulation (Fig. 2.3b) without the need to propose that the animal 'knows when it does not know' or any other metacognitive process.

Applications of Smith and colleagues' (2008) model depicted in Fig. 2.3a have broad implications for a variety of metacognition experiments. It is important to note that the model generates predictions that are *stimulus independent* in contrast to the traditional stimulus-response hypothesis. According to a stimulus-response hypothesis, an animal is assumed to learn to do a particular response in the presence of a particular stimulus. For example, with a stimulus-response mechanism, an animal can learn to select the decline response in particular stimulus conditions at a higher rate than in other stimulus conditions. A stimulus-response hypothesis has the form of an inverted U-shaped function in Fig. 2.3a for the decline response (i.e. in contrast to the constant attractiveness proposed by Smith et al.'s threshold in Fig. 2.3a). By contrast, according to a stimulus-independent hypothesis, previous reinforcement with a particular response is sufficient to produce that response in the future at a relatively low frequency. Because the response has a constant attractiveness, its use is independent of stimulus context. Although many studies in comparative metacognition are well equipped to test stimulus-response hypotheses, they are not adequate to test a stimulus-independent hypothesis.

Conflicting views on standards

Agreement on the standards by which evidence for metacognition is evaluated appears to be absent. In some domains of research, the standards by which evidence for a phenomenon are evaluated are set implicitly. For example, as initial evidence for a new phenomenon emerges, criteria may be set and refined by subsequent empirical or conceptual developments. In other domains, standards are set by the development of explicit, quantitative models, and advances may occur by pitting multiple models against one another.

Research in comparative metacognition has always been framed by application of principles of Morgan's canon (Morgan 1906). Yet application of Morgan's canon is, to some degree, subjective and open to multiple interpretations (Thomas 1998). To validate an animal model of metacognition, it is necessary to exclude more parsimonious alternative accounts. Inherently, validation of an animal model of metacognition is accomplished by exclusion (i.e. ruling out alternative

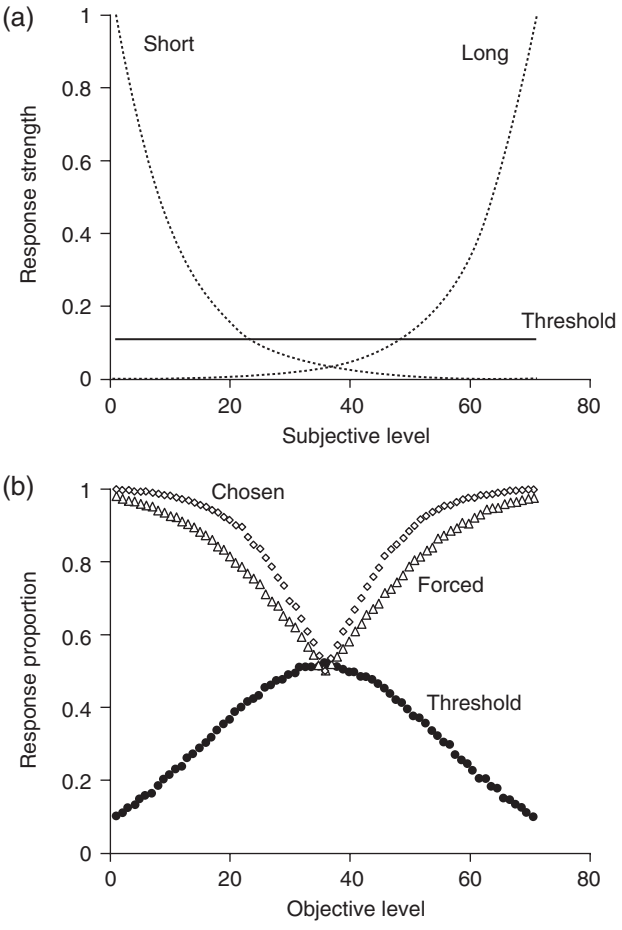


Fig. 2.3 Schematic of a response-strength model and simulation. a) Presentation of a stimulus gives rise to a subjective level or impression of that stimulus. For any given subjective level, each response has a hypothetical response strength. The schematic outlines response strengths for two primary responses in a two-alternative forced-choice procedure and for a third (i.e. decline) response (labelled threshold). Note that response strength is constant for the third response (i.e. it is stimulus independent). By contrast, response strength is highest for the easiest problems (i.e. at the extreme subjective levels). For the most difficult problems (i.e. middle subjective levels) the decline-response strength is higher than the other response strengths. With kind permission from Springer Science+Business Media: *Psychonomic Bulletin & Review*, The comparative study of metacognition: Sharper paradigms, safer inferences, 15(4), 2008, 679–91, J. David Smith, Michael J. Beran, Justin J. Couchman, and Mariana V. C. Coutinho. b) Simulation of schematic shown in (a). Simulation of a response-strength model with a flat threshold produces apparently functional use of the decline response (i.e. difficult stimuli are declined more frequently than easier stimuli). The choice–forced performance advantage emerged as a function of stimulus difficulty. With kind permission from Springer Science+Business Media: *Psychonomic Bulletin & Review*, The comparative study of metacognition: Sharper paradigms, safer inferences, 15(4), 2008, 679–91, J. David Smith, Michael J. Beran, Justin J. Couchman, and Mariana V. C. Coutinho.

accounts of the data). Validation of methods to document metacognition is an essential step, and it is important that researchers do not skip this essential step.

In the sections that follow, I outline a number of perspectives that may be applied to the evaluation of putative metacognition data. The perspectives that follow are not developments of criteria. Rather they are examples of navigating the interface between Morgan's canon and a growing body of research. With each new empirical milestone, a perspective on Morgan's canon can be offered. Perhaps the accumulation of challenges to a simple interpretation to putative metacognition data strengthens the confidence in a metacognition hypothesis.

Case studies

As the body of research on comparative metacognition has grown, it has become possible to apply a case-study approach. According to this perspective, a number of species have been evaluated for evidence of metacognition. Some species fail to show evidence for metacognition, and others provide putative metacognition data. According to a basic learning perspective, fundamental principles of learning may explain putative metacognition data. From the case-study perspective, one may wonder why some species fail and yet others provide putative metacognition data. Importantly, many types of species presumably rely on fundamental mechanisms of learning. Perhaps the difference between species can be explained by the presence or absence of metacognition—according to this perspective, species that provide putative metacognition data have metacognition, and other species do not. I refer to this perspective as a case-study approach because it involves drawing inferences from data obtained from a selection of available species. For example, a number of studies suggest that pigeons do not show metacognition (Inman and Shettleworth 1999; Sutton and Shettleworth 2008; Roberts et al. 2009; Adams and Santi 2011). Yet, rhesus monkeys have passed a variety of tests for metacognition (Smith 2009; Terrace and Son 2009), including some tests that are quite similar to those that pigeons failed (cf. Hampton 2001; Sutton and Shettleworth 2008). Along similar lines, a thorough effort was made to obtain metacognition data from capuchin monkeys (Paukner et al. 2006; Basile et al. 2009; Beran et al. 2009; Fujita 2009; Beran and Smith 2011). The capuchins failed to provide consistent evidence for metacognition despite the readiness of rhesus monkeys to provide putative metacognition data, often using the same methods (e.g. Beran and Smith 2011). Because rhesus and capuchin monkeys presumably both have basic mechanisms of learning at their disposal, perhaps the evidence for rhesus metacognition is strengthened by the absence of similar evidence from capuchins.

A number of concerns limit the conclusiveness of a case-study approach. It is possible that capuchins did not comprehend some aspect of the task in the same way as rhesus monkeys. For example, any difference in the immediacy of reward may contribute to different levels of performance. In the study by Beran and colleagues (2009), it may be argued that the uncertain response produced a delayed reward, whereas a middle response produced immediate reward. Because immediate reward is likely better than delayed reward, the species difference may derive from reward differences in the two types of tasks rather than from differential treatment of the stimulus continuum. Moreover, any difference in other cognitive or behavioural traits may complicate the case-study interpretation outlined earlier. For example, attention to the experiment or experimental contingencies may differ across species. Importantly, differences in impulsiveness, motivation, or perception may interact with experimental designs used to study metacognition; for example, the urge to seek out a food reward may affect how deliberative an animal is in choosing carefully among available response options. More broadly, it is always difficult to draw conclusions from negative evidence. The absence of evidence for metacognition is not evidence of absence of the capacity. Although this well-trodden inferential principle is not a basis to assert

that capuchins have a capacity for which no data exist, it should make us cautious to use capuchin data to interpret rhesus data.

Critical data patterns

Some measures of metacognition may provide stronger evidence than other empirical data patterns. For example, declining a difficult problem can be explained by learning to decline in the specific stimulus conditions encountered during training (i.e. stimulus-response hypothesis—in the presence of a particular stimulus, do a specific response). By contrast, a number of studies have assumed that other data patterns are more uniquely predicted by metacognition. For example, Inman and Shettleworth (1999) introduced the idea that it is critical to assess accuracy with and without the opportunity to decline difficult tests. They argued that an animal without metacognition would have the same level of accuracy when tested with and without the opportunity to decline tests. However, simulations conducted by Smith and colleagues (2008) suggest that a non-metacognition proposal can produce the same pattern of data.

Sophisticated materials

It is possible to search the available database for evidence of metacognition. A wide variety of approaches have been used and an even wider variety of stimuli have been used in studies of metacognition. It is possible that some evidence provides stronger support for metacognition than other evidence. For example, metacognition tasks have sometimes used perceptual judgements, application of concept formation, trial unique (or daily-trial unique) stimuli, same–different judgements, and memory tasks. Perhaps some of these approaches, in principle, provide stronger evidence for metacognition. According to this perspective, it is noteworthy that perceptual judgements may be more grounded in the stimulus conditions established during initial training. By contrast, memory tasks and other abstract judgements may be less grounded in stimulus conditions, more abstract, or more sophisticated. Hence, the sophistication of the task or the stimulus materials may be a factor to weigh in the evaluation of putative metacognition data.

A number of concerns limit the impact of sophistication of task or stimulus materials. An evaluation of the equivalent task or stimulus material outside of a metacognition context raises some concern about the applicability of this issue to weighing evidence for metacognition. Take, for example, the case of a memory task, for which putative evidence of metamemory in animals has been obtained (see Fig. 2.1 for procedure and data). Perhaps metamemory is more convincing than metacognition (i.e. outside the context of memory). This perspective is intuitive because memories are internal, and an appraisal of an internal memory would appear to provide direct access to an assessment of internal knowledge states. However, the sense in which memories are internal is shared with any representational account. Hence, an important issue to evaluate is whether a metamemory (i.e. metarepresentational) account is needed beyond a representational account. According to a representational account, the presentation of a stimulus gives rise to an internal representation of that stimulus; in the context of memory research, the internal representation decays with the passage of time after stimulus presentation. The strength of the dynamic representation likely determines a number of variables, such as accuracy, interference, competition with other stimuli, etc. Hence, forgetting, attention, and proactive interference may be studied from a representational perspective. It is not necessary to introduce a metamemory account to explain these basic cognitive operations because the representational account is adequate. An adequacy test can be applied to putative evidence for metacognition. If a fading trace account is sufficient to explain putative metacognition data, then little room is left for supporting a metarepresentational account. A fading trace account can be applied to familiar stimuli and to trial-unique stimuli.

A representational account of memory is needed, and the evaluation of metamemory requires an assessment of how much can be explained by a representational account before a metarepresentation account is offered. Next, a representational account (without proposing any metarepresentations) is offered for memory experiments that have been interpreted as evidence for metamemory.

The Smith et al. (2008) model may be applied to the case of metamemory by using a trace-decay continuum for a fading stimulus trace (Fig. 2.4). Thus, the model is sufficient to explain the decline rate and the chosen–forced performance advantage. It is important to evaluate the possibility that the monkey's performance depicted in Fig. 2.1 could be based on a *primary* representation of trace strength. According to this view, use of the decline response is based on a decaying memory trace just as the old–new responses from the primary task are based on a decaying memory trace. Because the same decaying memory trace (i.e. the same primary representation) is used for both the primary memory task and the decline response, a secondary representation may not be needed to explain the data.

Perhaps the use of two different responses (decline and matching responses) indicates that the two responses are based on different types of representations. The interpretive problem here is how to determine if the monkey is responding based on a primary representation (e.g. a weak stimulus representation) or based on a secondary representation (i.e. the monkey knows that it does not know the correct answer). It is noteworthy that in the case of human metacognition, behavioural data may be augmented by reports about subjective experiences which are not available for animals. Perhaps it is sufficient to claim that any paradigm that uses memory as the primary task will, by definition, result in secondary representations about memory, thereby providing evidence for metamemory. However, this perspective is problematic. What data specifically implicate the use of a secondary representation? Before Smith and colleagues (2008)

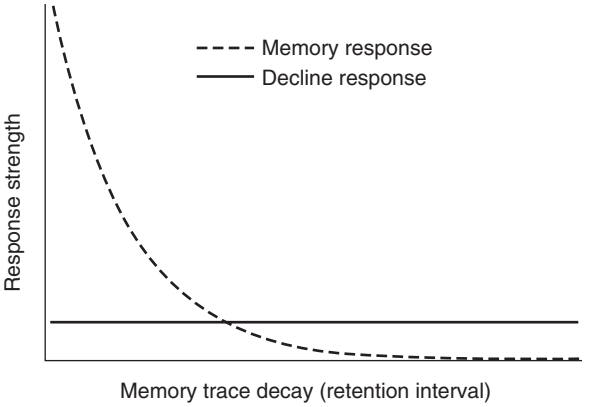


Fig. 2.4 Schematic of a response-strength model of metamemory. Presentation of a stimulus gives rise to a decaying memory trace. Trace decay (shown on the horizontal axis) grows as a function of retention interval. Response strength (i.e. a low-frequency threshold) is constant for the decline response. By contrast, memory response strength is highest for the shortest retention intervals. For the most difficult problems (i.e. long retention intervals), the decline response strength is higher than the memory response strength. The horizontal axis may be viewed as a primary representation (see text for details). From Crystal, J. D., and Foote, A. L. (2009). Metacognition in animals. *Comparative Cognition & Behavior Reviews*, 4, 1–16. © Crystal, J. D. Reprinted with permission.

documented putative metacognitive data using a non-metacognition model, the answer to this question was that the chosen–forced performance advantage could not be explained without appeal to metacognition. However, this pattern of data does not require a metacognition explanation. The burden of proof, in this situation, is on providing evidence that implicates a secondary representation, and until such evidence is provided the cautious interpretation is to claim that a primary representation is sufficient to explain the data. Moreover, the observation that the memory trace is an *internal* representation is not adequate to answer the question posed previously. Indeed, all representations are internal. If all that is needed is an internal representation (i.e. a primary representation), then what is to prevent the assertion that performance on matching to sample is based on metacognition (i.e. a secondary representation)? Perhaps the use of multiple responses in these types of experiment (i.e. the decline response and the primary response of choosing a correct/incorrect choice in matching to sample) strengthens a metamemory account. Accordingly, it may be argued that the non-decline responses are dedicated to reporting about the primary representation, whereas the decline response is dedicated to reporting about a secondary representation. The interpretive problem is that we do not know if this is the case. Clearly, an independent line of evidence is needed. In any case, Smith et al.'s model deals with competition between responses by selecting between responses based on low-level mechanisms without application of a secondary representation.

Hampton's (2001) study had several other elegant features which may strengthen a metacognition account. For example, after training with one retention interval, monkeys received no-sample probes to directly manipulate memory. It is intuitive that an animal with metamemory would respond adaptively by declining the test, which is what the monkeys did. However, this could also be based on a primary representation. Indeed, if a sample is omitted on a probe trial, the trace strength from the most recently presented sample (i.e. the stimulus presented on the trial that preceded the probe) would have an unusually long time to decay. Hence, the trace strength from the primary representation would be quite low and likely lower than the threshold for declining the test. Thus, a decline response would be expected based on a primary representation. How do we determine if the monkey is responding based on a primary representation (i.e. a weak primary stimulus representation) or based on a secondary representation (i.e. the monkey knows that it does not know the correct answer, in this case because there is no correct answer)?

Critical experimental techniques

Perhaps there are some empirical techniques that can provide evidence for metacognition that cannot be explained by non-metacognition proposals. A basic approach towards developing critical experimental techniques focuses on the role of extensive training with previous stimulus conditions. For example, it is possible that an animal learns to decline difficult problems because it detects that reward has been maximized in the past by choosing to decline in these stimulus conditions. A number of approaches have been used to test the reward-history alternative explanation. These approaches include trial-unique stimuli, transfer tests, omitting direct reward of uncertainty responses, and delayed feedback; perhaps one or more of these approaches is a critical experimental technique. Use of trial-unique stimuli limits the role of reward applied to the stimulus, but it does not limit the role of reward applied to the response. For example, in a recognition memory task, a subject is asked to judge a stimulus as new or old. Although the stimulus may be unique each trial, the new versus old response option (or stimuli associated with these response options) is constant throughout the experiment. Hence, a response strength is expected to accrue to the response options, and a winner-take-all response rule would favour the decline response when it is higher than old or new response options.

A transfer test is designed to test a stimulus response account of metacognition. The rational is as follows. If an animal has learned to select the decline response in a specific set of stimulus conditions, deprive the animal of these stimulus conditions to determine if it can flexibly decline in novel stimulus conditions (i.e. conditions in which it has not yet learned to use the decline response). However, the model developed by Smith and colleagues (2008) is stimulus independent. The model predicts the use of a decline response without learning to do so in specific stimulus conditions.

A number of studies have attempted to limit the role of direct reinforcement by not explicitly rewarding the decline response. One approach to this end involves moving to the next trial when a test is declined without providing any primary reinforcement (Smith et al. 2006). A more recent approach involves concealing the role of reinforcement on the primary task (e.g. line length discrimination) in addition to not directly rewarding the decline response (Beran et al. 2006). However, delay to reinforcement is a potent reinforcement variable (Kaufman and Baron 1968; Carlson 1970; Richardson and Baron 2008). Importantly, in each of these types of experiments a judicious selection of a decline threshold will maximize the number of reinforcements per unit time, and this can be accomplished without proposing metacognition (Crystal and Foote 2009). To examine the role of delay to reinforcement in these types of experiments, Crystal and Foote conducted a simulation of reinforcement rate. In the simulation, we used the feedback described by Beran et al. (2006). In the primary task, a correct response produced one food pellet, and an incorrect response did not produce any food pellets. Importantly, in their procedure, an incorrect response produced a timeout of 20 sec. An uncertainty response did not produce food and did not produce a time out. We used a flat uncertainty threshold, as proposed by Smith et al. (2008). In the simulations, we examined response strengths for the uncertainty response that varied from 0 to 1. The amount of food per unit time will be constant as a function of the threshold values in the simulations if delay to reinforcement is not a reward variable in these studies. By contrast, a particular threshold value for the uncertainty response will maximize food per unit time if delay to reinforcement functions as a reward variable.

Fig. 2.5 shows that there was a peak in food per unit time in the simulation. Thus, a subject in these types of experiments could adjust its threshold to maximize food per unit time, and this adjustment of the 'non-reinforced' uncertainty response is reinforced by reduced delay to reinforcement in the overall procedure. The simulation showed that despite the lack of direct reward of the uncertainty response, there are residual reinforcement variables at work in these types of experiments. Thus, the uncertainty response was indirectly reinforced by increased food rate; application of the Smith et al. (2008) model would predict use of the uncertainty response for the intermediate stimuli. Concurrent reinforcement may maintain the tendency to select the uncertainty response at a low frequency.

Most demonstrations of metacognition focus on the selection of a decline response more frequently in conditions of *high* task difficulty. Hampton (2001) provided a rare example of putative metacognition data using unusually *low* task difficulty. The monkeys were trained with a medium-length retention interval, and an unusually short retention interval was used as a transfer test. Perhaps the demonstration of unusually good performance is a critical test for metacognition. However, a response-strength account (Fig. 2.4) also predicts reduced use of a decline response in unusually easy tests. Hence, these data cannot be a decisive test.

An approach towards resolution of the status of comparative metacognition

The field of comparative metacognition has engaged in a productive debate about standards (Smith et al. 2003, 2009; Crystal and Foote 2009, 2011; Hampton 2009). Yet, progress in the

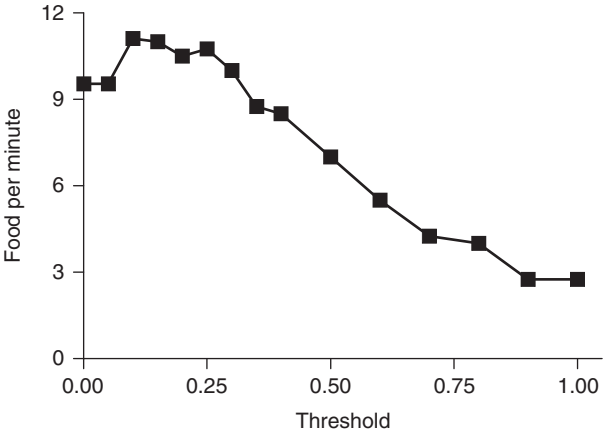


Fig. 2.5 Results of a simulation of reinforcement density as a function of variation in threshold for the uncertainty response using the generalization and constant-threshold concepts from Smith et al. (2008). Reinforcement and delay values were based on Beran et al. (2006). Although no food was delivered upon selecting the uncertainty response, the simulation showed that the value of the threshold for selecting the uncertainty response influenced the amount of food obtained per unit time in the primary discrimination. Thus, the uncertainty response was indirectly reinforced despite efforts to eliminate reinforcement. From Crystal, J. D., and Foote, A. L. (2009). Metacognition in animals. *Comparative Cognition & Behavior Reviews*, 4, 1–16. © Crystal, J. D. Reprinted with permission.

comparative analysis of metacognition may be threatened by conflicting views about the standards required to document metacognition in animals. The previous sections discussed experiments that may be interpreted as evidence for metacognition in animals, but these data may also be interpreted from a basic learning perspective. How can the debate be resolved? Standards for evaluating evidence of metacognition are needed. One approach centres on disputes about parsimony. Accordingly, metacognition may be interpreted as a relatively simple explanation. Indeed, metacognition in animals may not entail all aspects of metacognition in people. Moreover, metacognition is a single proposal, whereas learning explanations are frequently multifaceted (e.g. including generalization, response selection, etc.). From this perspective, metacognition may be viewed as the discrimination of an internal stimulus, which may not be fundamentally different from the discrimination of an external stimulus.

I suspect that debates about parsimony are not likely to lead to a resolution on the status of metacognition in animals. The ability to reflect on one's own mental processes is a defining feature of human existence. Consequently, determining whether animals have knowledge of their own cognitive states is a fundamental question. From this perspective, metacognition in animals is new and likely more complex than basic learning mechanisms. Moreover, a standard for metacognition should not be set too low, such that putative metacognition data are also readily explained by well-established basic principles of learning. A standard for metacognition in animals should be balanced. For example, it is not productive to set the standard so high that metacognition in animals cannot be demonstrated in any situation. An example of such a problem focuses on the view that metacognition is not well specified except by exclusion. Exclusion is a valuable approach, but it is important to exclude basic learning mechanisms.

An approach towards resolving the debate focuses on the use of well-specified models of both metacognition and non-metacognition hypotheses. The models should be specified at a level of

detail that can be simulated with computational procedures that match empirical procedures. Predictions of each model could be developed in detail. One virtue of this approach is that intuitions about predictions may not be correct. Simulations of well-specified models allow for correction of seemingly intuitive predictions, when required. Moreover, multiple versions of metacognition and non-metacognition models can be explored to evaluate the impact of various assumptions in a particular model. Some assumptions may be essential to generate a predicted pattern of data, whereas other features of a model might be robust over alternative assumptions. Simulations can be used to test existing behavioural methods. Moreover, refinements may be applied to develop new behavioural methods. To resolve the debate about comparative metacognition, behavioural methods that produce divergent predictions when applied to metacognition and non-metacognition models are needed. Empirical tests with such methods using various species will ultimately resolve the debate about metacognition in animals.

Acknowledgement

Supported by National Institute of Mental Health grant R01MH080052.

References

Adams, A. and Santi, A. (2011). Pigeons exhibit higher accuracy for chosen memory tests than for forced memory tests in duration matching-to-sample. *Learning & Behavior*, 39, 1–11.

Anderson, R. B. (2001). The power law as an emergent property. *Memory and Cognition*, 29, 1061–8.

Basile, B. M., Hampton, R. R., Suomi, S. J., and Murray, E. A. (2009). An assessment of memory awareness in tufted capuchin monkeys (*Cebus apella*). *Animal Cognition*, 12, 169–80.

Beran, M. J. and Smith, J. D. (2011). Information seeking by rhesus monkeys (*Macaca mulatta*) and capuchin monkeys (*Cebus apella*). *Cognition*, 120, 90–105.

Beran, M. J., Smith, J. D., Redford, J. S., and Washburn, D. A. (2006). Rhesus macaques (*Macaca mulatta*) monitor uncertainty during numerosity judgments. *Journal of Experimental Psychology: Animal Behavior Processes*, 32, 111–19.

Beran, M. J., Smith, J. D., Coutinho, M. V. C., Couchman, J. C., and Boomer, J. (2009). The psychological organization of 'uncertainty' responses and 'middle' responses: A dissociation in capuchin monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 371–81.

Carlson, J. G. (1970). Delay of primary reinforcement in effects of two forms of response-contingent time-out. *Journal of Comparative and Physiological Psychology*, 70, 148–53.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language*, 23, 58–89.

Church, R. M. (1997). Quantitative models of animal learning and cognition. *Journal of Experimental Psychology: Animal Behavior Processes*, 23, 379–89.

Crystal, J. D. and Foote, A. L. (2009). Metacognition in animals. *Comparative Cognition & Behavior Reviews*, 4, 1–16.

Crystal, J. D. and Foote, A. L. (2011). Evaluating information-seeking approaches to metacognition. *Current Zoology*, 57, 531–42.

Dunlosky, J. and Bjork, R. A. (Eds.) (2008). *Handbook of metamemory and memory*. New York: Psychology Press.

Foote, A. L. and Crystal, J. D. (2007). Metacognition in the rat. *Current Biology*, 17, 551–5.

Fujita, K. (2009). Metamemory in tufted capuchin monkeys (*Cebus apella*). *Animal Cognition*, 12, 575–85.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 5359–62.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comparative Cognition & Behavior Reviews*, 4, 17–28.

Inman, A. and Shettleworth, S. J. (1999). Detecting metamemory in nonverbal subjects: A test with pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 25, 389–95.

Kaufman, A. and Baron, A. (1968). Suppression of behavior by timeout punishment when suppression results in loss of positive reinforcement. *Journal of the Experimental Analysis of Behavior*, 11, 595–607.

Killeen, P. R. (2001). Writing and overwriting short-term memory. *Psychonomic Bulletin & Review*, 8, 18–43.

Morgan, C. L. (1906). *An introduction to comparative psychology*. London: W. Scott.

Paukner, A., Anderson, J., and Fujita, K. (2006). Redundant food searches by capuchin monkeys (*Cebus apella*): a failure of metacognition? *Animal Cognition*, 9, 110–17.

Richardson, J. V. and Baron, A. (2008). Avoidance of timeout from response-independent food: Effects of delivery rate and quality. *Journal of the Experimental Analysis of Behavior*, 89, 169–81.

Roberts, W. A., Feeney, M. C., McMillan, N., MacPherson, K., Musolino, E., and Petter, M. (2009). Do pigeons (*Columba livia*) study for a test? *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 129–42.

Sargisson, R. J. and White, K. G. (2001). Generalization of delayed matching-to-sample following training at different delays. *Journal of the Experimental Analysis of Behavior*, 75, 1–14.

Sargisson, R. J. and White, K. G. (2003). The effect of reinforcer delays on the form of the forgetting function. *Journal of the Experimental Analysis of Behavior*, 80, 77–94.

Sargisson, R. J. and White, K. G. (2007). Remembering as discrimination in delayed matching to sample: Discriminability and bias. *Learning & Behavior*, 35, 177–83.

Shepard, R. N. (1961). Application of a trace model to the retention of information in a recognition task. *Psychometrika*, 26, 185–203.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317–23.

Sikström, S. (1999). Power function forgetting curves as an emergent property of biologically plausible neural network models. *International Journal of Psychology*, 34, 460–4.

Smith, J. D. (2009). The study of animal metacognition. *Trends in Cognitive Sciences*, 13, 389–96.

Smith, J. D., Shields, W. E., and Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26, 317–73.

Smith, J. D., Beran, M. J., Redford, J. S., and Washburn, D. A. (2006). Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General*, 135, 282–97.

Smith J. D., Beran, M. J., Coutinho, M. V. C., and Couchman, J. C. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review*, 15, 679–91.

Smith, J. D., Beran, M. J., Couchman, J. J., Coutinho, M. V. C., and Boomer, J. B. (2009). Animal metacognition: Problems and prospects. *Comparative Cognition & Behavior Reviews*, 4, 40–53.

Sutton, J. E. and Shettleworth, S. J. (2008). Memory without awareness: Pigeons do not show metamemory in delayed matching to sample. *Journal of Experimental Psychology: Animal Behavior Processes*, 34, 266–82.

Terrace, H. S. and Son, L. K. (2009). Comparative metacognition. *Current Opinion in Neurobiology*, 19, 67–74.

Thomas, R. K. (1998). Lloyd Morgan's Canon. In G. Greenberg and M. M. Haraway (Eds.) *Comparative Psychology: A Handbook*, pp. 156–63. New York: Garland.

White, K. G. (2001). Forgetting functions. *Animal Learning & Behavior*, 29, 193–207.

White, K. G. (2002). Psychophysics of remembering: The discrimination hypothesis. *Current Directions in Psychological Science*, 11, 141–5.

Wixted, J. T. (2004). On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia *Psychological Review*, 111, 864–79.

Chapter 3

Are birds metacognitive?

Kazuo Fujita, Noriyuki Nakamura,
Sumie Iwasaki, and Sota Watanabe

Studies have shown that birds perform extremely well in various cognitive tasks despite their small brains. For instance, pigeons, a most well-studied avian species, form various concepts such as humans, trees, fish, artificial objects, paintings of a particular school, Baroque music, etc. (Herrnstein and Loveland 1964; Lubow 1974; Herrnstein et al. 1976; Herrnstein and de Villiers 1980; Porter and Neuringer 1984; Watanabe et al. 1995). They also discriminate same from different (Young et al. 1997; Young and Wasserman 2001), study prospectively in memory tasks (Roitblat 1980), and may even show 'insight' in problem-solving tasks (Epstein et al. 1984).

Cognitive animals may not necessarily be metacognitive. Inman and Shettleworth (1999) were the first to ask whether birds may be metacognitive also. They trained pigeons to peck at the same comparison figure that appeared earlier as a sample. One-third of the trials were simply regular delayed matching-to-sample trials as a memory test. On another one-third of the trials, another figure appeared alone instead of the comparisons after the delay period and the pigeons had no option but to peck at this escape stimulus. The last one-third of trials were the critical test trials, in which comparisons appeared with the escape stimulus. On these 'combined' trials, pigeons had a chance to choose between matching-to-sample to collect a large reward upon success, and pecking the escape key to collect a smaller but guaranteed reward. As a result, pigeons' matching accuracy was higher on combined trials than on regular forced matching trials. This was the test later called the 'concurrent metacognition' test (Terrace and Son 2009).

Might the pigeons have metacognitively judged whether they surely remembered the sample or not? Inman and Shettleworth (1999) tested the same birds on a slightly different procedure; now the birds had to choose to go to memory tests or to escape before they saw the comparison stimuli. This was the test later called the 'prospective metacognition' test (Terrace and Son 2009). In this test, however, no pigeons showed higher accuracy when they chose to take the memory tests than when they were forced to do so. No evidence for metacognitive judgement was obtained. This was in contrast with the performance by rhesus macaques (Hampton 2001) and capuchin monkeys (Fujita 2009) tested in similar procedures.

Sutton and Shettleworth (2008) replicated their earlier work in more varied procedures. In this study, pigeons not only failed in the prospective metacognition test but also in the easier concurrent metacognition test. They failed, too, in a retrospective metacognitive task, in which they chose between high-risk and low-risk options after performing a memory test. Similar failure was reported by Sole et al. (2003) using a conditional density discrimination task. Roberts et al. (2009) also reported that pigeons would not seek unseen samples in delayed matching tasks.

Despite such repeated failure, it is too early to conclude that pigeons are never metacognitive; we suspected that conditional discrimination tasks such as matching-to-sample tasks might be too demanding for pigeons to leave room for metacognitive processing within their cognitive resource. Conceivably, if working memory resources are scarce, due to the cognitive demands by

the primary task, there should be little room to do more cognitive tasks including metacognitive judgements.

Here we report two studies in pigeons and bantam chickens in simpler tasks. In one study, the birds performed on visual search tasks. In the other, pigeons performed on a simultaneous response chaining task successfully conducted by Terrace and colleagues (Straub et al. 1979; Straub and Terrace 1981) in this species.

Study 1: confidence judgements after visual search

Previous studies have demonstrated that rhesus macaques may be aware of how confident they are of their discriminated responses in perceptual and cognitive tasks (Son and Kornell 2005; Kornell et al. 2007). Similar retrospective metacognition was tested using a matching-to-sample task in pigeons but the result was negative (Sutton and Shettleworth 2008). In this study (Nakamura et al. 2011), we asked whether pigeons and bantam chickens would differentially choose icons that lead to either a larger reward only upon successful search or a smaller but guaranteed reward.

Training choice of confidence icons

Our subjects in Study 1 were six homing pigeons (*Columba livia*) and three bantam chickens (*Gallus gallus domesticus*). They had various laboratory experience of visual discrimination in the computer-controlled apparatus. They were kept at approximately 85–90% of their free-feeding body weights. All of the birds worked in the standard operant chamber with a touch-sensitive liquid crystal display (LCD) installed behind the opening on one wall.

The basic task was a visual search of a predetermined target and three homogeneous distracters. In experiment 1, the target was purple and the distracters were six colours surrounding this target that ranged in hue from very similar to the purple target to considerably different from the target (Fig. 3.1a). We first trained birds to peck at the target accurately in more than 60% of the trials for one session.

After the birds mastered this basic task, they were trained to peck at one of two icons (referred to as confidence icons) that appeared after the birds pecked at one of the four stimuli, either correctly or incorrectly. One icon was labelled the 'risk' icon. A peck at it after a correct peck at the target was reinforced by mixed grain accompanied by a flash of the food cup and the same response after an incorrect peck at a distracter was followed by a timeout. The other icon was labelled the 'safe' icon. A peck at it was reinforced by mixed grain accompanied by a flash with a probability of 33.3% or conditionally reinforced by a brief flash of the food cup with 66.7% probability regardless of the result of preceding visual search. At first only one confidence icon, either the 'risk' or 'safe' icon, was presented at the same frequency to familiarize the birds to the contingencies. Later, both confidence icons were presented in half of the trials and one of the confidence icons appeared with the same frequency in the other half of the trials (Fig. 3.2). Twelve sessions of 192 trials were conducted.

We compared the proportion of choice for the 'safe' icon after incorrect visual search (designated as 'I' in Fig. 3.3) with that after correct visual search ('C' in Fig. 3.3). Five of the six pigeons chose 'safe' more often after incorrect than after correct visual search. One of the pigeons and all three bantams failed to show this tendency, and they were retrained with a new set of colours (Set 2 in Fig. 3.1b). Two of the bantams still failed, and they were retrained with the first set of colours (Set 1). One of the bantams retired due to low motivation at this stage. As a result, all pigeons and two bantams learned to choose 'safe' more often after incorrect than after correct visual search, either after training with one or two sets of colours.

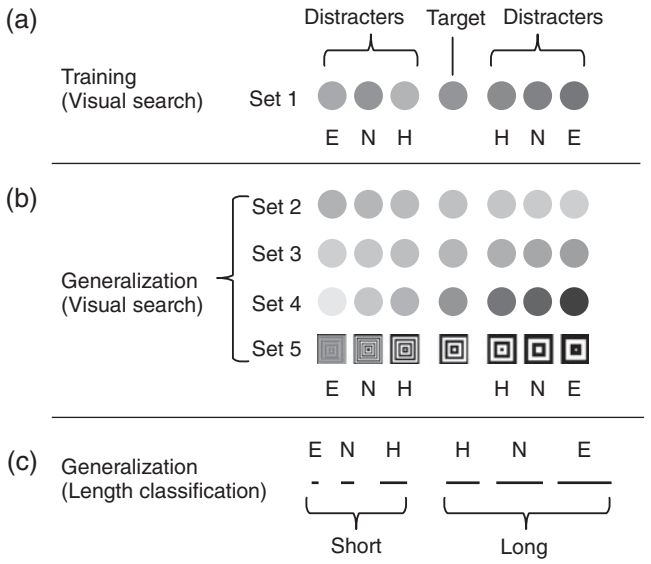


Fig. 3.1 Stimuli used in Study 1. a) Stimuli used in the training of visual search. b) Stimuli used to test generalization to visual search of new items. c) Stimuli used to test generalization to bar-length discrimination. E, N, and H denote easy, normal, and hard discrimination, respectively. With kind permission from Springer Science+Business Media: *Animal Cognition*. Do birds (pigeons and bantams) know how confident they are of their perceptual decisions?, 14(1), 2011, 83–93, Nakamura, N., Watanabe, S., Betsuyaku, T., and Fujita, K. (See also Colour Plate 1.)

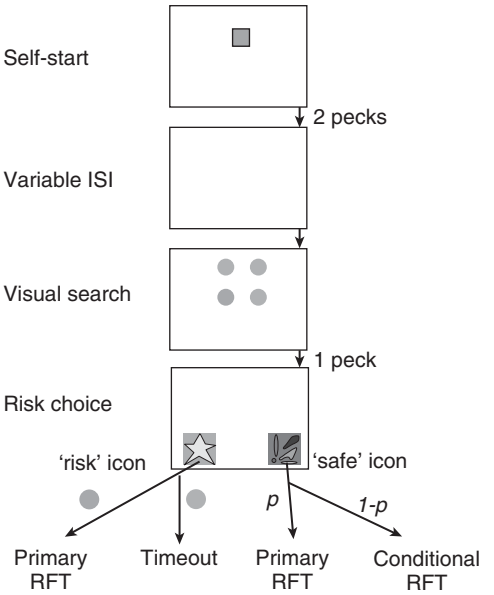


Fig. 3.2 A schematic diagram of the visual search task with risk choice. With kind permission from Springer Science+Business Media: *Animal Cognition*. Do birds (pigeons and bantams) know how confident they are of their perceptual decisions?, 14(1), 2011, 83–93, Nakamura, N., Watanabe, S., Betsuyaku, T., and Fujita, K. (See also Colour Plate 2.)

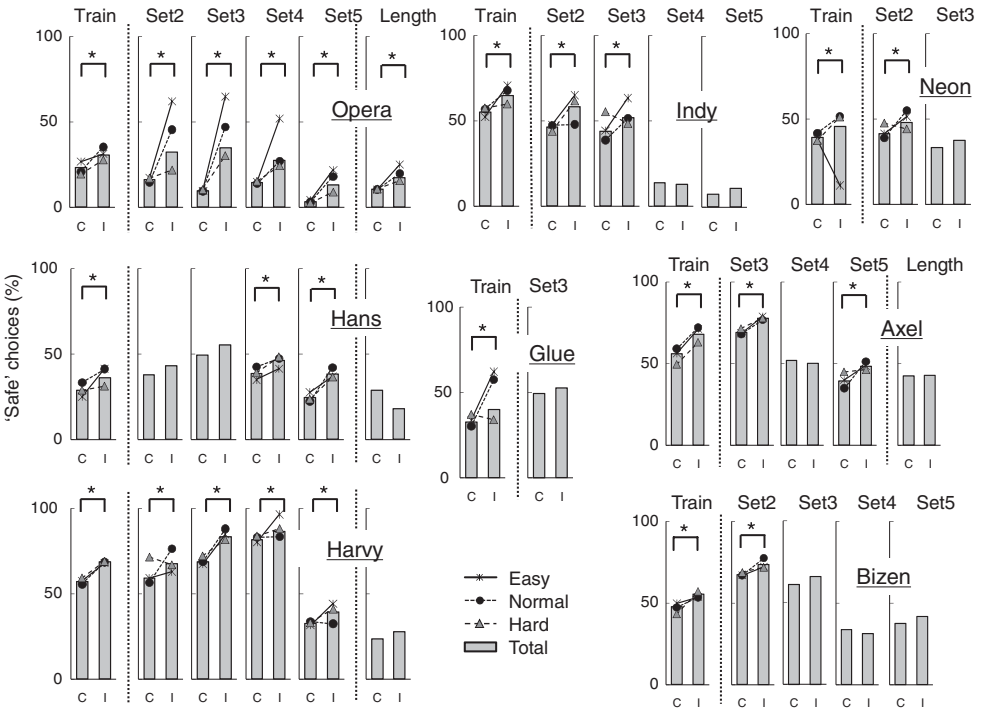


Fig. 3.3 Proportion of choice of 'safe' confidence icons after correct (C) and incorrect (I) responses, for individual birds. Opera, Hans, Harvy, Indy, Glue, and Neon are pigeons and Axel and Bizen are bantam chickens. Lines with symbols are for easy, normal, and hard discriminations and histograms are for the average of the three. With kind permission from Springer Science+Business Media: *Animal Cognition*. Do birds (pigeons and bantams) know how confident they are of their perceptual decisions?, 14(1), 2011, 83–93, Nakamura, N., Watanabe, S., Betsuyaku, T., and Fujita, K.

The higher frequency of 'safe' responses after an error was consistent with the idea that birds metacognitively recognized their confidence of visual search performances. However, such behaviour can, of course, be shaped by a simple operant conditioning through differential reinforcement in the final situation. Therefore we tested generalization of their choice for the 'risk' and 'safe' icons to new discriminations.

Generalization to new visual search tasks

We conducted a series of generalization tests using new stimuli listed in Fig. 3.1b for successful birds. Two sets were of different colours, one set was of greys of different brightness, and one set was of concentric squares of different thickness.

Prior to each generalization test, visual search in each set of stimuli was trained without confidence icons. Each generalization test was conducted in the same procedure as the last phase of previous training. Each test was conducted for 12 sessions immediately after the birds showed above 60% accuracy for 1 session.

Fig. 3.3 shows the results of six pigeons and two bantams. Different birds completed different numbers of generalization tests depending upon their generalization performances. However, as is clear from the figure, the proportion of choices of 'safe' icons was in general higher after

incorrect visual search (I) than after correct search (C) and in many tests the difference reached statistical significance. Most impressively, two pigeons (Harvy and Opera) showed this difference in all of the generalization tests. One pigeon (Hans) and one bantam (Axel) were also successful in most of the tests.

These good generalization performances were again consistent with the metacognition account, but another obvious possible account was that the birds may have simply learned the contingencies among the stimuli, their choice, and outcomes of the risk/safe choices in each of the tests. In order to examine this possibility, we calculated the disparity (referred to as ΔS) between the percentage of 'safe' choices after correct trials from that after incorrect trials in each session, and compared this score between the first six sessions and the last six sessions for each successful subject. There were no significant differences for any of the birds. Thus, the differential choice of confidence icons remained stable during the test sessions.

Another possibility may be that the birds based their icon choices not on their metacognitive judgement but on the apparent difficulty (shown by the stimuli themselves) of visual search. If the birds had done so, the proportion of 'safe' choices would have been a simple function of the difficulty of the task irrespective of the outcome of the visual search, correct or incorrect. This is clearly not the case; in Fig. 3.3, there were no consistent difference in the 'safe' choice proportion among easy, normal, and hard discriminations designated as E, N, and H in Fig. 3.1 (refer to Nakamura et al. (2011), for a more detailed discussion).

Still another possibility may be that the birds might have based their icon choices on their reaction time in visual search. The analyses of latency before pecking at one of the four visual search stimuli, however, found no reliable difference between correct and incorrect responses.

In sum, these additional analyses strengthened the interpretation that the birds were making confidence judgements of their preceding visual search performance. However, there may be other potential clues specific to the situation of the test. To test this possibility, we next examined generalization of their confidence icon choice to a completely different discrimination task, which was a bar-length classification task.

Generalization to a bar-length classification task

The birds who were successful in the last generalization test with concentric squares (Harvy, Opera, Hans (pigeons), and Axel (a bantam)) participated in this final test. All of them had experienced this basic task in search of their illusory perception (e.g. Nakamura et al. 2006, 2008, 2009).

The basic discrimination task went as follows; after an intertrial interval, a horizontal bar appeared on the display. There were six different lengths (Fig. 3.1c). Two or three pecks at this bar resulted in two response locations, one designated as 'short' and the other as 'long.' A peck at 'short' was correct for three short bars and a peck at 'long' was correct for three long bars. The birds were first trained on this basic task using bars from 10–80 pixels in length. After reaching 80% or higher correct classification, they went into the generalization test with confidence icons.

In this test, just like the ones described earlier, one or two confidence icons were presented after the classification response of bar length. Other conditions were the same; namely, half were one-icon (either 'safe' or 'risk') trials and the other half were two-icon trials and 12 sessions of 192 trials were conducted.

The results are shown in Fig. 3.3. One pigeon (Opera) chose 'safe' after incorrect trials significantly more often than after correct trial, as in previous tests on visual search. All of the other birds (Harvy, Hans, and Axel) failed to show a reliably differentiated choice of confidence icon choices between correct and incorrect bar-length classification responses.

We compared ΔS values described earlier (the percentage of 'safe' choices after incorrect trials minus that after correct trials) between the first six sessions and the last six sessions for Opera, but no significant difference was found. More importantly, the value of ΔS was positive (10.8%) in the very first session of this test. That means that this bird clearly generalized his confidence icon choice to this completely new task from the beginning. Further, the choice of the 'safe' icon after incorrect responses was not accounted for by the difficulty of the task (E, N, and H).

Discussion

Pigeons and bantams first learned to choose 'safe' and 'risk' icons after performing visual search responses with colours as stimuli. Most birds generalized this learning to visual searches with variously different colours, disks of different brightness, and concentric squares. Further, most impressively, one pigeon successfully generalized his choice of confidence icons to a completely different discrimination task, bar-length classification. His behaviour was comparable to rhesus macaques' similarly tested using other types of discrimination tasks (Kornell et al. 2007).

These results suggest that at least some pigeons and bantams may be capable of metacognitively recognizing the confidence of their discriminated operant responses in the immediate past. Many of the potential non-metacognitive clues were unlikely to control their behaviour; we have shown that difficulty of discrimination, gradual learning in each test, and reaction time cannot account for the birds' confidence icon choices.

Previous failure by pigeons in various metacognitive tasks may be, at least in part, due to a difficulty of the basic task. As discussed earlier, difficult discrimination tasks may leave little room in the birds' working memory to process metacognitive information. There still remains the possibility that the birds utilized some unidentified non-metacognitive clues available in the task. For instance, the birds may have chosen the 'safe' icon when they were simply distracted by some external noise, or when they simply did not look at the stimuli well. Although these might be usable for some of the trials, we do not think that these factors by themselves could give rise to the reliable, consistent, and long-lasting differentiated responses shown by our birds in this study.

Study 2: hint seeking in simultaneous chaining

When non-human primates have insufficient knowledge to solve a discrimination task, they sometimes will seek additional information. For instance, Call and Carpenter (2001) and Call (2010) showed that great apes looked into baiting tubes more often when they had not seen the experimenter hiding food in one of the tubes than when they had. Rhesus macaques performed similarly (Hampton et al. 2004). In a different situation, Kornell et al. (2007) showed that rhesus macaques sought hints for a next response in a simultaneous chaining paradigm, or list learning task, more often when they had started to learn a novel list than when they had mastered the list.

Although non-human primates perform metacognitively in these situations, pigeons again have been shown to fail in such situations. Roberts et al. (2009) gave pigeons the opportunity to either view a sample stimulus before they performed a matching-to-sample task or instead just perform the matching component without having seen the sample. Needless to say, matching to sample is impossible without knowing the sample of the trial, and so viewing the sample would have been the only correct choice. However, they chose between those options non-selectively, one leading to a sample followed by comparisons and the other leading directly to comparisons.

However, once again, matching to sample is not always an easy task for pigeons. In the present study we used a simultaneous chaining procedure to test concurrent metacognition in pigeons. In brief, we tested whether pigeons would seek a 'hint' for the next response more often when they did not know the correct sequence than when they had learned it. In other words, we wanted to

know whether there would be a negative correlation between their chaining accuracy and their frequency of 'hint' seeking (Iwasaki et al. 2010).

Learning set of simultaneous chaining

We first trained four pigeons to form a learning set of simultaneous chaining of three-item lists. They had various laboratory experiences with a touch monitor. One of them (Neon) participated in our previous study described earlier. The same operant boxes as in the previous study were used. By approximation, all of the pigeons were trained to peck at three items presented at a time in a predetermined sequence to receive mixed grain as a reward. A peck at a correct stimulus extinguished it at first. Next the same response flickered the pecked item for 3 sec. That is, the pecked item came back after 3 sec. This flicker time was gradually shortened to 100 msec at the final stage. A peck at a wrong item in any location in the sequence resulted in a timeout.

We repeated this training for 5–8 lists with novel computer icons as items. Finally the pigeons formed a set to learn a new list at the accuracy above 70% within 12 sessions of 60 trials.

Training to peck at a marked item and a 'hint' icon

After the birds formed the necessary learning set, they were trained to peck at an item that was marked by a white frame in the simultaneous chain. The white frame appeared either before the first or the second response in two-thirds of the trials. Training continued until the birds pecked at the marked item reliably at above 90% of the trials.

Then the birds were trained to peck at an icon that was going to be used as a 'hint' icon; in half of the trials, the 'hint' icon appeared alone and the birds simply pecked at it to receive a reward. In the other half of the trials, the ongoing regular simultaneous chain task appeared without the 'hint' icon. This was conducted for one session.

Simultaneous chaining with a 'hint' icon

In this stage, the birds performed a regular simultaneous chaining task in half of the trials (forced trials). In the remaining half of the trials, the 'hint' icon was presented with the three items of the list. The items for the chaining were novel; that is, the birds had to learn a new list.

In the trials with the 'hint' option, a peck at the 'hint' icon resulted in a flickering white frame on the item to be pecked at next. Up to three 'hint' requests, i.e. one for each item in the list, were possible within a trial. See Fig. 3.4 for a schematic representation of the procedure.

The same list was used until the birds reached 70% accuracy in the forced trials for two consecutive sessions. No more than 12 sessions for each list were given irrespective of the birds' accuracy.

The reward was the same irrespective of the use of the 'hint' icon for three of the birds. Because one bird (Clara) requested 'hints' in all of the trials with the 'hint' option, the probability of primary reinforcement in trials with the 'hint' claimed was lowered to 50% for this bird. For the remaining 50%, only the brief flash of the food cup was given as a secondary reinforcer. However, this resulted in this bird completely avoiding any requests for a 'hint', and the probability of primary reinforcement was raised to 75% in the 39th session. After he recovered using the 'hint' request in the 52nd session with the 11th list, the probability was kept at 75% and his data were collected from these final conditions (52nd session and afterwards). Three birds (George, Clara, and Neon) worked for six novel lists and Roki worked for five lists.

Results and discussion

Fig. 3.5 shows the correlation of the birds' accuracy of the chaining in the forced trials with no 'hint' option available and the proportion of trials in which the birds pecked at the 'hint' icon at least once within the trial. Each dot denotes a single session.

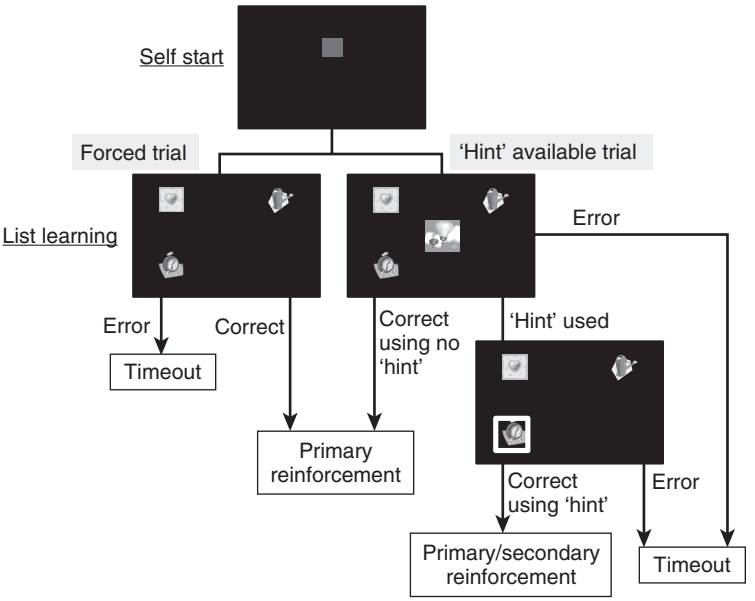


Fig. 3.4 A schematic diagram of the simultaneous chaining task with 'hint' option. (See also Colour Plate 3.)

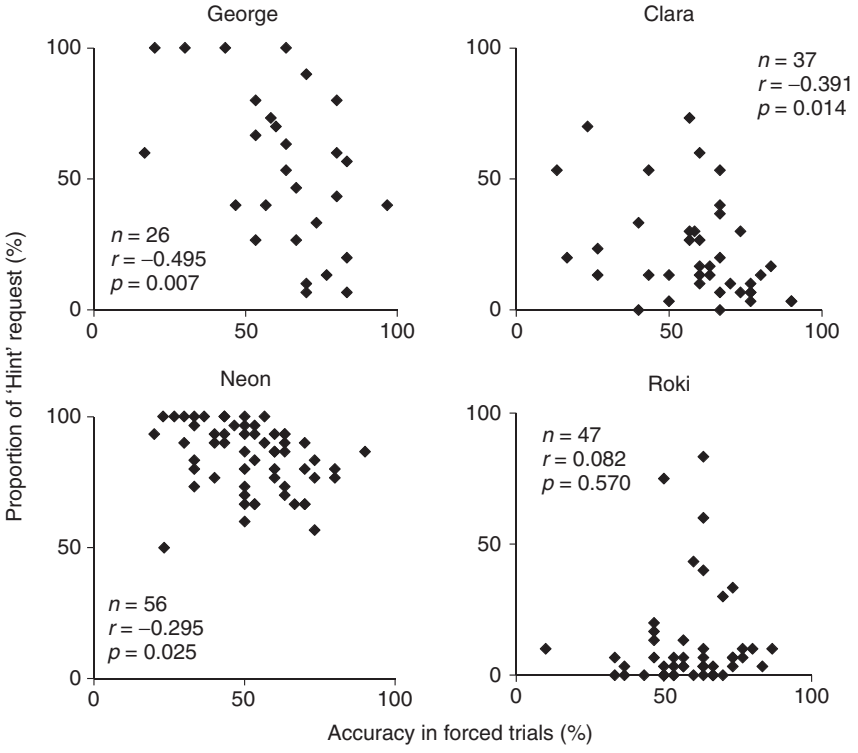


Fig. 3.5 Proportion of trials in which the birds requested a 'hint' at least once in hint-available trials as a function of the accuracy of the simultaneous chaining in the forced trials for individual pigeons. Each symbol denotes a session.

As shown, three of the four birds produced significantly negative correlations. That means they sought 'hints' more often when their chaining accuracy was lower than when it was higher. We additionally compared the proportion of trials in which the 'hint' was requested between the first session and the last session for each list for the successful birds. This analysis revealed significant differences for George and Clara but not for Neon. Therefore, those two birds sought 'hints' more often at the initial stages of learning each list than at the final stages.

These results were consistent with the view that some of our pigeons may have metacognitively recognized their knowledge of correct sequence of each list, or how well they had learned the sequence, and used this recognition to adaptively control their behaviour when 'hint' was available upon request. There are obvious other non-metacognitive possibilities. For instance, the pigeons may have learned to choose the 'hint' icon more often when they found stimulus items unfamiliar. However, we think that this is unlikely because we used no more than six lists after they had acquired use of the 'hint' icon and it does not immediately seem easy for the birds to learn this kind of strategy from such meagre opportunities. Otherwise, they may have pecked at the 'hint' icon whenever they hesitated to choose the next item in the list. Or, as pecking at the 'hint' icon and pecking at a next item compete, the birds may have been more likely to peck the 'hint' icon whenever their tendency to choose a next item got weak (response competition; see Hampton 2009).

Possibilities of the birds' use of these external cues may be tested by examining generalization of their apparently metacognitive behaviour to new situations. For George and Clara, we changed the basic task from simultaneous chaining to visual search with coloured disks. In half of the trials, there appeared the same 'hint' icon next to the visual search display, which, if pecked, could show the correct target. We prepared easy trials and difficult trials by changing the similarity of the distracters to the target. Although the birds' visual search accuracy changed as a function of this difficulty, we found no negative correlation of search accuracy and 'hint' seeking responses. Thus, we failed to reject the possibility for some potential non-metacognitive clues included in the original simultaneous chaining task, although such rejection does not mean we can conclude that pigeons are metacognitive.

Discussion

We have shown two potentially metacognitive performances in birds. In Study 1, pigeons and bantams adaptively selected two icons after performing visual search. That is, they chose the 'safe' icon more often after incorrect search than after correct search. This suggests the possibility that pigeons and bantams retrospectively judged their confidence in their preceding discriminated operant responses. Additional analyses rejected several other non-metacognitive clues that could have controlled the birds' choices and, most impressively, one pigeon generalized his choice of the confidence icons to a completely different discrimination task, a bar-length classification. Such findings seem persuasive that some birds are capable of retrospective metacognition, although we have yet to assess other unidentified metacognitive clues as a potential explanation.

In Study 2, we tested whether pigeons asked for 'hints' whenever they lacked good knowledge about what they should do next. In brief, pigeons learned to peck at three items in a predetermined sequence as a list. Whenever they started to learn a new list, they had to find a correct sequence by trial and error. In half of the trials, there was a 'hint' icon; pecking at it showed the next item to choose. Three of the four pigeons pecked at the 'hint' icon more often when their accuracy on the sequence learning was low than when it was high; that is, there was a negative correlation between frequency of 'hint' requests and the accuracy of list learning. This suggests pigeons' metacognitive judgement about their knowledge of the list. However, their 'hint' requesting did

not generalize to a completely new task that involved visual search. Thus the evidence for this concurrent metacognition is still weak. There are possibilities that other non-metacognitive factors may have controlled the pigeons' 'hint' requesting.

In view of the previous failures in demonstrating birds' metacognitive performance (e.g. Sutton and Shettleworth 2008), our successful results, though not yet conclusive, imply that the cognitive demand necessary to solve a task difficult for birds may keep the birds from learning to utilize metacognitive information they could access in cases. Neither of our basic tasks requires a conditional discrimination in which the birds have to process hierarchical information. The seemingly metacognitive performance shown by pigeons and bantams in Study 1 appears a quite general one that could apply to similar tasks in which stimulus dimensions to attend are varied and, in one pigeon, a more complicated conditional position discrimination of bar length even was performed. This may mean that once the birds learned to use their metacognitive clues in simpler tasks, they could generalize this learning to more complicated situations. This may apply at least for retrospective metacognition.

In contrast, performance by pigeons in Study 2 seems less general. It may have been something specific to the situation where the birds learned their use of the 'hint' icon, or simply it may not have been strong enough. Notwithstanding, the positive results for the birds' use of metacognitive judgement of their knowledge, or learning stage, within one type of discrimination task shed some light on what avians may be endowed with regarding this complicated introspective cognitive processing.

There may be various other non-metacognitive accounts for the performances of birds in the current series of study. For instance, Smith et al. (2008) discussed that reinforcing 'uncertainty' responses (e.g. use of risk icons and hint icons in the current studies) might lead the seemingly metacognitive behaviour to instead be the result of associatively motivated responses. Jozefowiez et al. (2009) also suggested seemingly metacognitive behaviour to be a result of economy in which the subjects simply learned it as beneficial. Although we acknowledge that these considerations are important, we believe that such increases in the reinforcement or benefits is in fact why an organism takes advantage of metacognition at all. We cannot expect use of metacognition to develop without any benefit. At the least, we believe that differential use of risk icons and hint icons depending upon the discrimination performances in the various generalization tests makes simple application of these hypotheses less likely. Clearly, however, more work on various aspects of avian metacognitive ability is required to better understand how this ability has evolved in the animal kingdom and to discuss how cognition and metacognition are interrelated.

Acknowledgements

Preparation of this manuscript was supported by the Grant-in-Aide for Scientific Research (S), No. 20220004 to KF, from the Japan Society for the Promotion of Sciences (JSPS) and by the Global COE Program, D-07 to Kyoto University, from Japan Ministry of Education, Culture, Sport, Science and Technology (MEXT).

References

Call, J. (2010). Do apes know that they could be wrong? *Animal Cognition*, 13, 689–700.

Call, J., and Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition*, 4, 207–20.

Epstein, R., Kirshnit, C., Lanza, R. P., and Rubin, L. (1984). 'Insight' in the pigeon: antecedents and determinants of an intelligent performance. *Nature*, 308, 61–2.

Fujita, K. (2009). Metamemory in tufted capuchin monkeys (*Cebus apella*). *Animal Cognition*, 12, 575–85.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 5359–62.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comparative Cognition & Behavior Reviews*, 4, 17–28.

Hampton, R. R., Zivin, A., and Murray, E. A. (2004). Rhesus monkeys (*Macaca mulatta*) discriminate between knowing and not knowing and collect information as needed before acting. *Animal Cognition*, 7, 239–46.

Herrnstein, R. J. and de Villiers, P. A. (1980). Fish as a natural category for people and pigeons. In G. H. Bower (Ed.) *The psychology of learning and motivation*, pp. 59–95. New York: Academic Press.

Herrnstein, R. J. and Loveland, D. H. (1964). Complex visual concept in the pigeon. *Science*, 146, 549–51.

Herrnstein, R. J., Loveland, D. H., and Cable, C. (1976). Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 2, 285–302.

Inman, A. and Shettleworth, S. J. (1999). Detecting metamemory in nonverbal subjects: A test with pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 25, 389–95.

Iwasaki, S., Watanabe, S., and Fujita, K. (2010). Hato ni okeru hinto-kikyu kadai wo mochiita meta-ninchi no kento. [A test of metacognition in pigeons using a hint-seeking task] Paper presented at the 70th Annual Meeting of Japan Society for Animal Psychology, Tokyo, Japan (in Japanese).

Jozefowiez, J., Staddon, J. E. R., and Cerutti, D. T. (2009). Metacognition in animals: how do we know that they know? *Comparative Cognition & Behavior Reviews*, 4, 29–39.

Kornell, N., Son, L. K., and Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18, 64–71.

Lubow, R. E. (1974). High-order concept formation in the pigeon. *Journal of the Experimental Analysis of Behavior*, 21, 475–83.

Nakamura, N., Fujita, K., Ushitani, T., and Miyata, H. (2006). Perception of the standard and the reversed Müller-Lyer figures in pigeons (*Columba livia*) and humans (*Homo sapiens*). *Journal of Comparative Psychology*, 120, 252–61.

Nakamura, N., Watanabe, S., and Fujita, K. (2008). Pigeons perceive the Ebbinghaus–Titchener circles as an assimilation illusion. *Journal of Experimental Psychology: Animal Behavior Processes*, 34(3), 375–87.

Nakamura, N., Watanabe, S., and Fujita, K. (2009). Further analysis of perception of the standard Müller-Lyer figures in pigeons (*Columba livia*) and humans (*Homo sapiens*): Effects of length of brackets. *Journal of Comparative Psychology*, 123(3), 287–94.

Nakamura, N., Watanabe, S., Betsuyaku, T., and Fujita, K. (2011). Do birds (pigeons and bantams) know how confident they are of their perceptual decisions? *Animal Cognition*, 14, 83–93.

Porter, D. and Neuringer, A. (1984). Music discriminations by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 10, 138–48.

Roberts, W. A., Feeney, M. C., McMillan, N., MacPherson, K., Musolino, E., and Petter, M. (2009). Do pigeons (*Columba livia*) study for a test? *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 129–42.

Roitblat, H. L. (1980). Codes and coding processes in pigeon short-term memory. *Animal Learning & Behavior*, 8, 341–51.

Smith, J. D., Beran, M. J., Couchman, J. J., and Coutinho, M. V. C. (2008). The comparative study of metacognition: sharper paradigms, safer inferences. *Psychonomic Bulletin and Review*, 15, 679–91.

Sole, L. M., Shettleworth, S. J., and Bennett, P. J. (2003). Uncertainty in pigeons. *Psychonomic Bulletin and Review*, 10, 738–45.

Son, L. K. and Kornell, N. (2005). Metacognitive judgments in rhesus macaques: Explicit versus implicit mechanisms. In H. Terrace and J. Metcalfe (Eds.) *The missing link in cognition: Origins of self-reflective consciousness*, pp. 296–320. New York: Oxford University Press.

Straub, R. O., Seidenberg, M. S., Bever, T. G., and Terrace, H. S. (1979). Serial learning in the pigeon. *Journal of the Experimental Analysis of Behavior*, 32, 137–48.

Straub, R. O. and Terrace, H. S. (1981). Generalization of serial learning in the pigeon. *Animal Learning & Behavior*, 9, 454–68.

Sutton, J. E. and Shettleworth, S. J. (2008). Memory without awareness: Pigeons do not show metamemory in delayed matching to sample. *Journal of Experimental Psychology: Animal Behavior Processes*, 34, 266–82.

Terrace, H. S. and Son, L. K. (2009). Comparative metacognition. *Current Opinion in Neurobiology*, 19, 67–74.

Watanabe, S., Sakamoto, J. and Wakita, M. (1995). Pigeons' discrimination of Monet and Picasso. *Journal of the Experimental Analysis of Behavior*, 36, 165–74.

Young, M. E. and Wasserman, E. A. (2001). Evidence for a conceptual account of same-different discrimination learning in the pigeon. *Psychonomic Bulletin & Review*, 8, 677–84.

Young, M. E., Wasserman, E. A. and Garner, K. L. (1997). Effects of number of items on the pigeons's discrimination of same from different visual displays. *Journal of Experimental Psychology: Animal Behavior Processes*, 23, 491–501.

Chapter 4

Seeking information in non-human animals: weaving a metacognitive web

Josep Call

Humans are all too familiar with metacognitive experiences. We all have experienced situations in which even though we cannot recall someone's name, we feel very confident that we will do so at a later time even without any external help. Many of us check that we have our passports and plane tickets before we leave the house for an overseas trip even though we remember fully well where we placed them the day before. Even the act of tying a string around one's finger is an act grounded in metacognition—anticipating the possibility that we may forget to do something we place an unusual mark on our bodies as a reminder.

These are just three examples of human metacognition but we could easily conjure many more examples that would attest to the central role that metacognition, defined as knowing about our own psychological states, plays in our everyday experiences. Indeed, metacognition plays such a central role that it is hard to disentangle it from other psychological processes, and it is tantalizing to think that metacognition may be the cognitive ability that separates humans from animals.

My goal in this chapter is not to focus on human metacognition. Instead, my goal will be to investigate the question of whether non-human animals (henceforth animals) possess metacognition. To do so, I will trace the emergence, development, and current state of one of the paradigms that has been used to investigate metacognition from a comparative perspective: the information-seeking paradigm. Once that is done, I will establish links between this paradigm and other paradigms that have been used to investigate metacognition in animals. Some of them have a relatively long history while others are incipient attempts to develop alternative approaches to this fascinating question. Note that the adjective 'alternative' is not meant to indicate 'replacement', but 'complement'. In this section, it will become apparent, much to my surprise I have to admit, that the various approaches are currently producing a quite coherent picture that I will try to capture in the next section. In the final section, I will propose two main research lines that may foster growth in this field.

Metacognitive origins

Over the last two decades, questions about the knowledge that non-human animals possess about their physical and social worlds has gained increasing research attention. Social knowledge, in particular, has received considerable attention. Researchers have assessed what non-human animals know about their social worlds both in terms of the social relations among their group members and the psychological states that they may infer to others. With regard to the latter, Premack and Woodruff's (1978) seminal paper on the attribution of intentions opened the way to questions of whether animals, most notably the great apes, attribute mental states to others. Although Premack and Woodruff (1978) contemplated attribution of mental states both to the self and others, most of the work on mental state attribution both in developmental and

comparative psychology has been devoted to investigate the mental states of others (see Call and Tomasello (2008) for a review). Only relatively recently, researchers have focused their attention on perhaps the more basic problem of what access individuals have about their own mental states.

Interestingly, the comparative study of the individual's own mental states, unlike the work aimed at investigating others' mental states, did not originally emerge from the mindreading tradition pioneered by Premack and Woodruff (1978) or the precursors that one can find in developmental psychology (e.g. Flavell 1978). Instead it emerged from cognitive psychology in the context of psychophysical research. Smith et al. (1995, 1997) pioneered a method to investigate whether subjects were able to make metacognitive judgements when presented with uncertain situations. In their original escape response paradigm, subjects were presented with a discrimination task in which the stimuli varied in difficulty. Like many discrimination tasks, subjects were presented a stimulus and offered a choice of two options (one correct and the other incorrect) but additionally, they were offered the possibility to decline trials. That is, by pressing a third option, subjects could skip the current trial and move to the next one. The analysis of these so-called escape responses showed that subjects preferentially declined trials presenting difficult discriminations. In other words, as uncertainty regarding the likelihood of success increased so did the frequency of declined trials. Such results were obtained in various species including rhesus macaques, humans, and dolphins and using a variety of discrimination tasks. Smith and colleagues concluded that subjects were capable of monitoring their uncertainty states.

One criticism that those initial studies received was that subjects associated the escape response with reinforcement, or put differently, that certain stimuli configurations acted as discriminative stimuli for selecting the escape response due to its reinforcing qualities relative to the penalties associated with those particular stimuli (see Smith et al. 2003). Note that these studies required extensive training for animals to learn to use the escape response properly. Consequently, subjects may not have been responding to their internal state of uncertainty but to the stimuli that they perceived. Additional studies, however, showed that the escape response was not simply tied to particular configurations because subjects were able to generalize the use of the escape response to novel discriminations that had not been associated with the escape response (e.g. Washburn et al. 2006). Moreover, the presentation of massed trials and delayed reinforcement considerably weakens this criticism, thus suggesting that individuals may indeed be responding to their perceived states of uncertainty (Couchman et al. 2010). This debate, however, is far from settled and advocates of the non-metacognitive account have proposed increasingly elaborate explanations to account for the continuous stream of studies reporting new findings based on new procedures (e.g. Crystal and Foote 2009). I think that it is perhaps fair to say that the field has entered a sort of arms race in which increasingly elaborated non-metacognitive explanations are met with ever more sophisticated empirical evidence which in turn generate increasingly more complex non-metacognitive explanations.

Another approach to comparative metacognition

Call and Carpenter (2001) took a different approach to the question of comparative metacognition. Unlike the approach of Smith and colleagues, Call and Carpenter's approach was grounded in the mindreading/developmental tradition. More specifically, the procedure that they devised relied heavily on the question of visual perspective taking and access to perceptual mental states. Instead of presenting individuals with ambiguous stimuli and training them to use an escape response, they presented subjects with incomplete information about the location of a reward in a foraging task and measured whether they sought additional information before they produced

a response. The main idea was whether subjects would be sensitive to their lack of information and remedy this situation before making a choice.

The original set-up of the information-seeking task was quite simple. Subjects faced two parallel tubes placed on a low lying platform with their openings oriented towards them. The experimenter placed a piece of food inside one of the tubes on the far side from where the subject was located. In order to obtain the food, all that the subject had to do was to touch the baited tube on their first attempt. No second choices were allowed. There were two conditions. In the visible condition, subjects witnessed the experimenter placing the food inside the tube whereas in the hidden condition the baiting took place behind an opaque occluder that blocked the subject's visual access. After the baiting was completed, the platform was pushed forward and the subject was allowed to select one of the tubes by touching it. Whether subjects looked inside the tubes before choosing was the main dependent measure and was never trained prior to the test.

Chimpanzees, orangutans, and 2½-year-old children looked inside the tubes before choosing more often in the hidden than the visible condition (Call and Carpenter 2001). Later studies using this same paradigm confirmed these results in other chimpanzees and orangutans and extended them to other species including gorillas, bonobos, and rhesus macaques (Hampton et al. 2004; Call 2005, 2010; Basile et al. 2009). In contrast, capuchin monkeys and dogs produced mixed results (Bräuer et al. 2004; Paukner et al. 2006). Call and Carpenter (2001) interpreted their initial result as an indication that subjects had access to their own mental states; in this case, subjects knew that they had seen or not seen the reward.

Three features of this task deserve special mention. First, from a practical point of view the implementation of this task is extremely simple. The apparatus is easy (and inexpensive) to build and the subjects require no special training since they only need to indicate their choices by touching one of the tubes—something that apes did quite spontaneously. The dependent measure of looking inside the tube, unlike the escape response, appeared spontaneously and it was not trained in any way. Second, and related to the previous point, the looking response is richer (and more open-ended) than the escape response since the former allows the experimenter to measure not only whether subjects produce the response, but also how they produce it. For instance, imagine that individuals A and B both look inside the tubes during a hidden trial but whereas individual A looks inside every tube possible (i.e. exhaustive search), individual B only looks inside the minimum number of tubes required to locate the reward (i.e. efficient search). Although both subjects produced the appropriate response, their different allocation of looks provides very useful information about the information controlling their search. Moreover, as we will see in the following sections, varying the features of the food hiding containers as well as the cost associated with looking (or not looking) can also reveal useful information that can play a major role in helping us narrow down the possible interpretations for the data available.

Third, from a theoretical point of view, this paradigm is closer to the area of mindreading than psychophysics. Several decades ago, Flavell (1978, 1993) established a connection between visual perspective-taking and metacognition research in children (see also Flavell et al. 1983). Such a connection did not exist in the comparative literature because researchers solely investigated whether chimpanzees and other animals knew what others could or could not see from different locations (see Call and Tomasello (2008) for a review). Recently, however, Krachun and Call (2009) have shown that chimpanzees also know where they have to position themselves to see a particular object. Thus, chimpanzees (and perhaps other animals as well) know both whether someone can see object A from their current location and they also know where they have to position themselves to see object A.

Alternative explanations

Although Call and Carpenter (2001) tentatively interpreted their initial data as being consistent with a metacognitive account, other interpretations are indeed possible, as these authors themselves acknowledged. Thus, the initial findings are also consistent with several non-metacognitive explanations. In general, there have been two sorts of alternative explanations that differed in the broadness of their explanatory scope. The first sort of alternative is composed of 'narrow-beam' explanations, typically aimed at accounting for one particular finding, but making no specific predictions about future findings. The second alternative is a 'broad-beam' explanation based on identifying a non-metacognitive construct that can potentially account for the observed results without recourse to the monitoring of knowledge states. Unlike the narrow-beam explanations, this other approach has the very desirable feature of having the potential to explain multiple findings and, more importantly, it can generate testable hypotheses that can guide future research. New data, however, have also challenged each of these two families of explanations. Next, we present the various alternatives that have been proposed and the findings that have weakened them.

Narrow-beam explanations

The generalized search hypothesis constituted the first non-metacognitive attempt to explain the original findings of the information-seeking paradigm (Call and Carpenter 2001; see also Carruthers 2008; Perner Chapter 6, this volume). According to this hypothesis, subjects that lack information about the location of the food engage in search behaviour until they locate the reward. Two features of this search are important. First, it is automatically triggered upon detecting an information deficiency and second, the search is unguided (i.e. random). In fact, the search behaviours of this sort would be used both to disambiguate information (e.g. a cat moving its head sideways to determine whether that thing is a mouse; Carruthers 2008) or simply when an individual lacks information as in the case of the hidden condition of the seeking information task (Perner Chapter 6, this volume). Once the reward is viewed (or the information is disambiguated), the search ceases and subjects attempt to retrieve it. This 'search, locate, and retrieve' routine is presumably something that many animals are equipped with and will commonly use when foraging for hidden items.

At face value, this hypothesis can explain the difference between visible and hidden trials in the original study. Additionally, this hypothesis can explain why subjects search more often with increasing delays after they have seen the location of the reward (Call 2010). Here, a delay between encoding the food location and food acquisition may have prevented subjects from retrieving that information, thus transforming an originally visible condition into a hidden condition in which subjects no longer can access their knowledge about the location of the food. However, there are three aspects to the observed data that do not fit with this explanation.

First, on a sizable percentage of trials which vary between 16% and 30% depending on the studies and the species, subjects selected the correct alternative after having inspected the tube that was empty (Call and Carpenter 2001; Call 2005). That is, subjects stopped searching before they actually viewed the reward, which does not fit with the searching until locating the food before choosing strategy. At the very least, this result indicates that subjects were able to disambiguate the situation without seeing the reward by using inference by exclusion (Premack and Premack 1994; Call 2004, 2006). This represents a more complex routine than the original search until locating the reward routine (see also Perner Chapter 6, this volume).

Second, contrary to what this hypothesis postulates, subjects' search is not random but guided. Krachun and Call (2009) presented chimpanzees with three types of containers: cylinders, triangles,

and trapezoids. Prior to the test, they allowed the subjects to explore those objects. Two of the objects were completely new and had never been associated with food. After the exploration was over, Krachun and Call (2009) placed a set of three identical containers on a platform forming a straight line. In different trials subjects faced either three cylinders placed on the platform in an upright position, three triangles, or three trapezoids and the experimenter baited one of the containers. The crucial aspect of this study was that owing to the containers' diverse geometry and their position on the platform, subjects had to position themselves in different locations depending on the container to spy the food. In particular, subjects had to climb and look from above to see the food inside the upright cylinders, they had to move to the side of the platform to see the triangle openings, and they had to move behind the trapezoids, thus changing their position 180 degrees with respect to their original location, to see the food inside them. Just like in previous studies, subjects looked inside the containers more often when they had not witnessed the baiting compared to when they had seen it. More importantly, the search was non-random. In particular, chimpanzees looked from above, side, and behind for the cylinders, triangles, and trapezoids, respectively. They did so even if only their first look was considered. This result is completely inconsistent with random search and demonstrated that subjects knew exactly where they had to position themselves to see the food.

Third, the generalized search hypothesis does not explain why subjects also look when they have witnessed the baiting. One could attribute those looks to the fact that they may have forgotten the location of the food, especially after longer delays. However, this does not explain why when they look, they do not look randomly but preferentially look inside the baited container, thus showing that they still remember the location of the food (Call 2010). Moreover, looks also exist even after short delays when the memories have not degraded yet, and again those looks (unlike those in the hidden condition) are preferentially targeted to the baited rather than the unbaited container (Call and Carpenter 2001). Another explanation is therefore necessary.

A second hypothesis that has been proposed to explain why subjects look inside tubes is because they may like the sight of the food (Perner pers. comm.). Here subjects would not be attempting to remedy some informational shortcoming, but seeing the reward per se has some hedonic value. Although this could certainly explain why they look inside tubes, it does not explain why they look more often in hidden than visible trials, why looking increases as a function of delay, why they select on about 25% of the trials after having not seen the reward or why once they have located the reward, they do not look again but choose. Moreover, if the tubes are shaken with the consequence that the baited tube produces a rattling sound, looking inside the tubes is reduced (Call 2010). Interestingly, the reduction is only observed in those subjects who had been able to use the noise made by the reward to infer its location in a previous study (Call 2004). Unless one postulates that hearing the reward also possesses hedonic value that replaces seeing the food, but this only occurs for those individuals capable of establishing a causal relation between the presence of the food and the production of noise, this result is hard to explain. Nevertheless, this hypothesis awaits empirical scrutiny.

Finally, the response competition hypothesis postulates that selecting one alternative will take precedence over looking unless the lack of information decreases the strength of the selection response (Hampton et al. 2004). One recent result, however, is completely inconsistent with this hypothesis. When subjects are presented with high-quality versus low-quality food, they look more often before choosing when the high-quality rather than the low-quality food is at stake (Call 2010). And they do so even if the baiting is conducted in full view of the subject. This is precisely the opposite result predicted by the response competition hypothesis since higher-quality food should reduce looking by increasing the strength of the reaching response. Interestingly, control tests showed that subjects remembered equally well the location of

high- and low-quality food. Call (2010) interpreted this result as an indication that subjects may entertain the possibility of being wrong in their choices and that is why they check more often when high-quality food is at stake even though their memories are equally good for both types of food.

Other studies have shown that individuals take cost into account when deciding whether they visually inspect inside the tubes before choosing. Hampton et al. (2004) found that increasing or decreasing the cost of looking into the tubes by placing the tubes lower or higher from the face of rhesus macaques affected the likelihood that they would look inside them. Call (2010) also investigated whether increasing the cost for looking affected the subjects' responses in gorillas, chimpanzees, and bonobos. However, instead of raising or lowering the tubes, he placed them on a fixed platform either in an oblique or a straight orientation with respect to the subject. This meant that looking inside oblique-oriented tubes was harder than looking inside straight-oriented tubes. Results confirmed those of Hampton showing that looking responses decreased as cost of looking increased. Interestingly, the reduction of looking was more pronounced when subjects had seen the baiting (42%) than when they had not witnessed it (13%). This suggests that subjects were more likely to forgo looking when they had already seen the reward.

Broad-beam explanations

Carruthers (2008) has taken a different approach in criticizing the studies on animal metacognition. Rather than postulating specific rules to explain specific results, as previous hypotheses had done, Carruthers (2008) offered a more complete non-metacognitive model to account for the evidence presented in metacognition studies. This model is based on postulating that animals, including humans, react to the level of anxiety that they perceive in certain situations. Thus, subjects are not reacting to their knowledge states but to the anxiety produced by those knowledge states. According to Carruthers (2008) an individual's knowledge states are opaque to the individual—something that also often applies to humans who may have the illusion that they know what they are reacting to when in reality they do not know.

One of the most appealing aspects of Carruthers' proposal is that an anxiety-mediated behaviour has a potential broad application. In fact, Carruthers (2008) used this model to strip metacognition out of the escape response and seeking information paradigms. Focusing on the seeking information paradigm, Carruthers' model can easily explain a variety of findings including the difference between visible and hidden conditions and the increase in looking as a function of delay since the baiting took place. It can even explain the so-called passport effect (Call and Carpenter 2001; Call 2010; i.e. the likelihood of seeking information increases directly proportional to the cost of failing to locate the reward) since the anticipation of not receiving the high-quality food may generate more anxiety than not receiving the low-quality food. However, there are some findings that this account cannot fully explain.

First and foremost, there is the search specificity. Recall that when subjects sought information, they engaged in targeted searches, they did not search randomly. The anxiety model, however, cannot explain this result because it postulates that the search for additional information is triggered automatically and gathered randomly. In this model, there is no room for directed searches because this would mean that subjects can distinguish between the information that they possess and the information that they are missing, and additionally, they would know how to remedy their informational shortcomings. Search specificity, however, can be explained by postulating that subjects must engage in some form of perspectival abilities about what they need to see to make the correct selection.

Second, there is the question of the interchangeability of information. Recall that subjects who heard the rattling of the reward inside the tube, looked inside the tube less than when they were

not offered this auditory information. Under the anxiety-based model, it is unclear why hearing the reward would ameliorate the anxiety (thus reducing looking responses) produced by not having seen the reward—unless the subjects treat seeing and hearing the food as equivalent from a point of view of their choices. Moreover, note that not all subjects treated this additional auditory information in this way. Only those who were able to use auditory information to infer the food location were able to reduce their looking behaviour in the metacognitive task. Again, it seems that anxiety alone would not have predicted this outcome especially since this only applies to those individuals capable of making inferences.

Building bridges between paradigms

So far we have mostly concentrated on the results of the information-seeking paradigm. Next we turn our attention to other paradigms that have been used to investigate metacognition to find out whether both sets of results are consistent with each other at various levels of analysis.

Delay and accuracy

Hampton (2001) observed that rhesus macaques increased their escape responses in a delayed matching to sample (DMTS) task as a function of the delay between the presentation of the sample and the alternatives—the longer the delay between the presentation of the sample and the selection of one of the alternatives, the more likely subjects were to select the escape option. This result is comparable to the data showing that subjects were more likely to seek information after longer delays (Call 2010). Kornell et al.'s (2007) study on hint seeking in macaques is also consistent with the data from these other tasks. Kornell et al. (2007) trained monkeys to touch a series of stimuli in a certain order. Touching the stimuli set in the correct order produced a reward whereas touching them in the wrong order produced no reward and a timeout. Once subjects had learned this basic task, they learned that they could press another key to request a hint about what was the next stimulus that needed to be touched to complete the correct sequence. Results showed that hint requests decreased as the monkeys' accuracy on the sequences increased. These data are complementary with the two previous studies that showed that escaping and information seeking also decreased with accuracy, which in turn decreased with time.

Risk and gambling

Suda-King (2008) pioneered a method that combines elements from both the escape response and the information-seeking paradigms. Suda-King (2008) investigated how orangutans responded to situations in which they had to choose between a 100% chance ($P = 1.00$) of getting a low-quality food or a variable probability (but always $P < 1.00$) of getting a high-quality food. The probability of getting the high-quality food was modified by varying the number of containers under which a single (high-quality) piece of food could be hidden. Just like in the information-seeking paradigm, Suda-King (2008) administered to subjects both visible and hidden trials. In visible trials subjects witnessed where the food was placed whereas in the hidden trials subjects were prevented from seeing the food's destination. Results showed that subjects were more likely to select the low-quality food in hidden trials. This is equivalent to selecting the escape response and netting the low-quality but certain reward in Hampton's (2001) study. Moreover, it shows that when facing incomplete information subjects not only seek additional information but also escape when they have no way to acquire information about the food location.

Haun et al. (2011) tested all great ape species in the Suda-King (2008) paradigm but varied both the size of the reward (equivalent to food quality in Suda-King's (2008) study) and the number of different containers where the reward could be hidden. Additionally, subjects received both

visible and hidden trials. Results showed that apes were more likely to select the smaller but safer reward in the hidden compared to the visible trials, corroborating Suda-King's (2008) results. Not all species, however, showed the same threshold for selecting the low-quality reward when facing a lack of information. In general, bonobos were more likely to select the smaller but safer option than chimpanzees and orangutans. These findings corroborated a previous study showing that bonobos were more risk averse than chimpanzees (Heilbronner et al. 2008). Moreover, all species reduced their choice of the smaller but safer reward as the size of the risky option was increased or the probability of finding it was increased by reducing the number of cups under which the risky option could be found. In fact, the number of cups available and the size of the reward under the cups explained more than 70% of the observed variance in the subjects' choices. Crucially, we found no evidence that repeated trial presentations affected the subjects' choices. In other words, the response patterns did not change within a session or between sessions. Thus, apes' sensitivity to the number of cups available and the size of the risky reward in hidden trials suggests that they can estimate, at least implicitly, the likelihood that their choices will be successful and then choose optimally in many cases.

The two previous studies possess another desirable feature in terms of metacognition. They involve the assessment of risk. Much work on human metacognition is based on asking subjects to make judgements about how certain they are about something or the likelihood that they will be able to recall some piece of information. Non-human animals cannot be directly interrogated in the same way as humans, but researchers have found ways to measure confidence in an indirect manner. This indirect method is based on the idea that subjects that are more confident about their responses may be more likely to take risks. One way to measure this is with the visible–hidden manipulation indicated previously, especially when it involves other factors such as the number of cups available where the high-quality food may be hidden. Recall that subjects are less willing to gamble when they have not seen the reward location in the risky choices.

Son and Kornell (2005) assessed risk in a different way in rhesus macaques—a way that makes it more similar to the way people are asked to make certainty judgements. Macaques were required to judge the length of a stimulus presented on a computer screen and once they had made a choice they could decide how much food they wanted to gamble in that particular trial before they were given feedback about their response. Son and Kornell (2005) observed that when facing an easy discrimination, monkeys gambled big (and won big) whereas when they faced a difficult discrimination they were more conservative and gambled smaller quantities. In a follow-up test, monkeys were able to transfer their 'gambling' skills previously associated with length discrimination to a task that involved assessing the pixel density of stimuli and another task that entailed recalling the picture of the object that they had been shown previously out of a collection of multiple objects.

Apes, macaques, and capuchins

The comparisons between species tested on different paradigms have also produced a rather coherent picture. Macaques have produced evidence consistent with metacognition both in the escape response paradigm (based both on discrimination and DMTS tests) and the information-seeking paradigm (e.g. Smith et al. 1997; Hampton 2001; Hampton et al. 2004). Similarly, apes have produced positive results both in the information seeking and the risk–safe reward paradigms (Call and Carpenter 2001; Suda-King 2008; Haun et al. 2011). Although Haun et al. (2011) found differences between bonobos and the other great ape species in the risk–safe task, Heilbronner et al. (2008) also found those interspecific differences even though their study did not assess metacognition. Thus, those interspecific differences may be related to risk-proneness rather than metacognition. All other studies that have included multiple great ape species have

found no differences among them (e.g. Call and Carpenter 2001; Call 2005, 2010). In contrast to the studies on macaques and apes, studies on capuchin monkeys have produced inconsistent evidence for metacognition in this species. In fact, mixed results for capuchin monkeys have been found in almost every paradigm tested including the seeking information and the escape response paradigms based on DMTS and other discrimination tasks previously used with macaques (Paukner et al. 2006; Basile et al. 2009; Beran et al. 2009; Fujita 2009).

As Smith (2009) has noted, the mixture of positive and negative results across species poses an interesting challenge to non-metacognitive explanations that invoke psychological processes that are shared by many species. In particular, it is unclear how the reinforcement history and anxiety postulated to explain the positive results in metacognition experiments (e.g. Carruthers 2008; Crystal and Foote 2009) can also explain the negative results of some species. After all, those species that fail the metacognitive tests do possess those psychological processes. Indeed, some of those experiments show that capuchins, just like macaques, can learn the tasks' basic requisites, but unlike macaques they do not seem to respond in the same way when facing uncertain situations (Paukner et al. 2006; Basile et al. 2009; Beran et al. 2009; Fujita 2009).

Appearance and reality

Recently, another research avenue based on the appearance-reality distinction has been assayed in non-human animals as a tool to investigate metacognition. Developmental psychologists have used this method extensively to investigate children's appreciation of their own and others' mental states. According to Flavell et al. (1983), recognizing and distinguishing appearance from reality is a commonplace experience for humans, one that is metacognitive in nature. Appearance-reality tasks typically involve presenting an object whose appearance leads children to judge its true nature incorrectly. For instance, children are presented with a sponge that looks like a rock and asked about it. When children judge it to be a rock, they are shown the true nature of the object. Researchers then ask children both about what the object *looks like* and what the object *really is*. Prior to 4 years of age, most children respond to both questions in the same way. In some cases, some children presented with an object behind a magnifying state that the object actually becomes smaller after the removal of a magnifying glass. In contrast, by the age of 4 these appearance reality errors have decreased dramatically.

Similar to the information-seeking and uncertainty-monitoring tasks, subjects in appearance reality tasks are confronted with ambiguous stimuli that create some perceptual/cognitive conflict that they need to resolve. This conflict, however, is not created by missing information or difficult discriminations but by contradictory information. Consequently, solving the task does not entail escaping or seeking additional information but being able to ignore appearances and focus on reality. Appearance-reality tasks in children typically involve presenting subjects with an object whose appearance leads children to judge its true nature incorrectly, something that is commonly seen in children before their fourth birthdays. Flavell and colleagues (1983) have argued that the difficulty for dealing with perceptual appearances is related to a more general limitation about analyzing the origin and properties of their mental representations.

The eminently verbal nature of the task has prevented researchers from using the same task with non-verbal organisms. Therefore, Krachun et al. (2009) used an indirect method to investigate how chimpanzees dealt with misleading appearances. In general, chimpanzees have a strong tendency to prefer larger over smaller grapes. However, when a smaller and a larger grape were placed behind a maximizing and a minimizing glass, respectively, thus reversing their virtual sizes, about 60% of the chimpanzees still selected the larger grape. They did this even when they were prevented from tracking the spatial position of the grape and they could only use the deceptive appearance of the grape to decide which one of the two grapes was larger. Moreover, two

control tests demonstrated that subjects did not solve this task by using a mechanism akin to that seen in reversed reward contingency tasks (i.e. pick the small grape to net the large one; Boysen and Berntson 1995). Therefore, this study showed that some chimpanzees were quite capable of overcoming appearances, which suggests that they, just like 3-year-old children, may possess some appreciation of the origin of their own perceptions and how the interposition of certain objects (e.g. magnifying glass) between themselves and the food affects them. It is still an open question whether chimpanzees can also maintain an awareness of misleading information once they know the true nature of an object. Additional research is needed to answer this question and to extend these findings to other appearance-reality tasks as well as to other species.

A synthetic view

It is certainly premature to attempt an integrated view of the field let alone trace the evolution of metacognition based on the available evidence. The data are too few both in terms of the species investigated and the number of paradigms that have been used in a relatively small number of individuals. So this will be by necessity an incomplete exercise but a necessary one as a way to assess where we stand and where we can go in the next few years. My main conclusion in this section is that the responses observed in metacognitive studies possess substantial *cognitive flexibility*. Both aspects are important. They are cognitive because not all available data can be reduced to anxiety monitoring, and they are flexible because individuals are capable of using information of various kinds, which in some cases includes making inferences, to produce efficient responses. In the next section, I will highlight two main research avenues through which the field may further develop.

How individuals respond to uncertainty has been the main question that research on metacognition in animals has investigated to date. Much less is known about other aspects of metacognition such as whether individuals are also capable of making confidence judgements in uncertain situations (but see Son and Kornell 2005). I draw this distinction because it seems to me that making confidence judgements is a particularly refined indication of uncertainty monitoring since subjects in such tasks not only have to detect their uncertainty but they also have to evaluate its level. It is uncontroversial that animals facing either ambiguous stimuli, or stimuli that create an internal cognitive conflict, behave differently than those who are not facing a conflict. In particular, individuals experiencing a cognitive conflict change their response latency, waver between competing responses, and even opt for responding to several options simultaneously. Research on metacognition has also documented that, when faced with uncertainty, individuals may escape the situation or seek additional information to resolve the uncertainty. Such responses may also be accompanied by changes in the individual's anxiety levels as indicated by several physiological (e.g. skin conductance) and behavioural (e.g. scratching) indicators.

Although much of the debate on metacognition has revolved around the notion that responses produced by individuals merely reflect the contingencies of reinforcement of the responses used to measure metacognition (see Crystal and Foote 2009; Smith et al. 2006; Smith 2009), another more stimulating debate has emerged regarding the nature of the indicators perceived by the uncertain individual (Carruthers 2008). More specifically, this debate revolves around the notion of whether subjects facing uncertainty merely detect those behavioural and physiological indicators of uncertainty noted earlier or they also take an extra step and also detect the causes of those indicators. Currently there are three pieces of evidence suggesting that some animals at least, may go beyond detecting the uncertainty indicators and may have some insight into the causes of those indicators.

First, when individuals are missing the information needed to choose correctly, they know where to look for it, and they engage in targeted searches. I will not elaborate further on this point

because it has been treated in the previous section. But briefly stated, detecting anxiety cannot predict where they will look because to do that, one has to know what one is looking for and what precisely needs to be done to find it. Second, individuals show a remarkable ability to integrate information from other modalities and thus replace the missing information appropriately. This is evidenced by the ability to replace missing visual information about the location of the food for the sound that the food makes when the baited tube is shaken. Note that only those individuals who can infer that the food is the cause of the auditory cue respond appropriately, which shows that the individual has to interpret the information in a causal manner. Even in cases in which the search produces no new information such as an auditory cue but just an empty tube, individuals can use this information to make an inference by exclusion and stop searching. Note that in both cases, spatiotemporal or cause–effect parameters play a crucial role in helping the individual infer the location of the food. Needless to say, in both cases once the sight of the food has been replaced by either a rattling sound or the sight of an empty tube, individuals correctly choose the baited tube.

Finally, when the search cost for the food is increased, there is some evidence indicating that animals can skip that search but mostly when they have seen the baiting, not when they did not witness it. If anxiety is regulating their looks, the decrease should occur in both conditions equally, but it does not. This result is reminiscent of animals' search strategies with low- and high-quality food. Even though they look more when high-quality food is at stake, they remember both types of food equally well. In other words, in visible trials both increasing the cost of searching and decreasing the cost of choosing incorrectly (by using a low-quality reward) produce the same effect: reduced looks. In such cases, searches become optional and individuals know it.

In summary, the great apes can resolve uncertainty by seeking information in a targeted manner, by replacing the missing information appropriately with equivalent information via inferential processes or forgoing the need to search for additional information when they have seen the food location and they still remember it. Taken together, these findings suggest that at least the great apes have some access to the causes of their uncertainty and they can deploy flexible means to remedy this situation. Given the data available on macaques in other metacognitive paradigms, it is likely that future studies will reveal that macaques' information seeking, just like that of the great apes, is targeted, integrated, and facultative.

Future directions

In the course of the last decade comparative metacognition has established itself as a fast growing research area. Obviously not all questions have been resolved and there is still a lively debate about whether data can be explained without recourse to metacognitive explanations. However, this debate should not obscure the fact that real progress has taken place. Not only is there now a rich body of evidence available, but many of the original non-metacognitive explanations once proposed are not tenable anymore in light of newer data. Rather than weakening the field, this back-and-forth debate between new data and new alternative explanations has created an invigorating effect, a likely indicator of continued growth in the next few years. I envisage two main research lines where this growth will take place.

The first research line is centred within each paradigm devoted to investigating metacognition. For instance, using the information-seeking paradigm will allow researchers to explore additional questions such as whether individuals can anticipate that they will forget some piece of information. Currently, all the work done to date has tested whether individuals can judge whether their information is still adequate to locate the reward. Will subjects that still have current access to this information be able to anticipate that it may degrade over time? One way to test this would be to

show that subjects 'study' longer or leave special marks when they know that the retrieval event will take longer (see Schneider and Sodian (1988) and Sodian and Schneider (1990) for related paradigms).

Still within the information-seeking paradigm, another direction in which the field can progress is by testing the predictions made by the anxiety-mediated theory. One first step in this endeavour will require a precise measurement of the levels of anxiety that individuals experience when they opt for seeking or not seeking information. A second step will consist of manipulating an individual's levels of anxiety either by behavioural or pharmacological means and observing the effects on information seeking. However, the application of certain manipulations (e.g. pharmacological) will have to be restricted to certain groups of individuals or species. Finally, one important question that deserves research attention is whether individuals who cannot currently recall a piece of information are nonetheless able to reliably predict that they will be able to recall that information in the future, the so-called tip-of-the-tongue phenomenon in humans. Such data would allow comparative researchers to strengthen the links with researchers focusing on human metacognition.

The second research line that should be developed further, especially if we are to trace the evolution of metacognition, consists of establishing stronger connections between paradigms within the comparative literature. One of the most encouraging signs that this may be a viable alternative is the convergence of findings from different paradigms and species that this chapter and others (e.g. Smith 2009) have highlighted. However, this tentative connection needs to be tested more systematically and rigorously. The next step would require using batteries of metacognitive tasks administered to the same individuals to see how they relate to each other. Crucially, each of the tasks forming the battery should include not only the basic task but also several of its variations that would allow researchers a more fine-grained assessment of an individual's metacognitive capabilities.

Furthermore, this method needs to be applied to multiple species so that one can begin to trace the evolution of metacognitive abilities. An added benefit of this increased connectivity between paradigms is that the two traditions in psychology that have guided metacognitive research in non-human animals, cognitive and developmental, may forge closer links with each other.

References

Basile, B. M., Hampton, R. R., Suomi, S. J., and Murray, E. A. (2009). An assessment of memory awareness in tufted capuchin monkeys (*Cebus apella*) *Animal Cognition*, 12, 169–80.

Beran, M. J., Smith, J. D., Coutinho, M. V. C., Couchman, J. J., and Boomer, J. (2009). The psychological organization of 'uncertainty' responses and 'middle' responses: A dissociation in capuchin monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 371–81.

Boysen, S. T. and Berntson, G. G. (1995). Responses to quantity: Perceptual versus cognitive mechanisms in chimpanzees (*Pan troglodytes*). *Journal of Experimental Psychology: Animal Behavior Processes*, 21, 82–6.

Bräuer, J., Call, J., and Tomasello, M. (2004). Visual perspective taking in dogs (*Canis familiaris*) in the presence of barriers. *Applied Animal Behaviour Science*, 88, 299–317.

Call, J. (2004). Inferences about the location of food in the great apes (*Pan paniscus, Pan troglodytes, Gorilla gorilla, Pongo pygmaeus*). *Journal of Comparative Psychology*, 118, 232–41.

Call, J. (2005). The self and the other: a missing link in comparative social cognition. In H. S. Terrace and J. Metcalfe (Eds.) *The missing link in cognition: origins of self-reflective consciousness*, pp. 321–341. New York: Oxford University Press.

Call, J. (2006). Inferences by exclusion in the great apes: the effect of age and species. *Animal Cognition*, 9, 393–403.

Call, J. (2010). Do apes know that they can be wrong? *Animal Cognition*, 13, 689–700.

Call, J. and Carpenter, M. (2001). Do chimpanzees and children know what they have seen? *Animal Cognition*, 4, 207–20.

Call, J. and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12, 187–92.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind and Language*, 23, 58–89.

Crystal, J. D., and Foote, A. L. (2009). Metacognition in animals. *Comparative Cognition and Behavior Reviews*, 4, 1–16.

Couchman, J. J., Coutinho, M. V. C., Beran, M. J., and Smith, J. D. (2010). Beyond stimulus cues and reinforcement signals: A new approach to animal metacognition. *Journal of Comparative Psychology*, 124, 356–68.

Flavell, J. H. (1978). The development of knowledge about visual perception. In C. B. Keasey (Ed.) *Nebraska symposium on motivation*, vol. 25, pp. 43–76. Lincoln, NE: University of Nebraska Press.

Flavell, J. H. (1993). The development of children's understanding of false belief and the appearance-reality distinction. *International Journal of Psychology*, 28, 595–604.

Flavell, J. H., Flavell, E. R., and Green, F. L. (1983). Development of the appearance-reality distinction. *Cognitive Psychology*, 15, 95–120.

Fujita, K. (2009). Metamemory in tufted capuchin monkeys (*Cebus apella*). *Animal Cognition*, 12, 575–85.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 5359–62.

Hampton, R. R., Zivin, A., and Murray, E. A. (2004). Rhesus monkeys (*Macaca mulatta*) discriminate between knowing and not knowing and collect information as needed before acting. *Animal Cognition*, 7, 239–54.

Haun, D. B. M., Nawroth, C., and Call, J. (2011). Great apes' risk-taking strategies in a decision making task. *PLoS ONE*, 6, e28801.

Heilbronner, S. R., Rosati, A. G., Stevens, J. R., Hare, B., and Hauser, M. D. (2008). A fruit in the hand or two in the bush? Divergent risk preferences in chimpanzees and bonobos. *Biology Letters*, 4, 246–9.

Kornell, N., Son, L., and Terrace, H. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18, 64–71.

Krachun, C. and Call, J. (2009). Chimpanzees (*Pan troglodytes*) know what can be seen from where. *Animal Cognition*, 12, 317–31.

Krachun, C., Call, J., and Tomasello, M. (2009). Can chimpanzees (*Pan troglodytes*) discriminate appearance from reality? *Cognition*, 112, 435–50.

Paukner, A., Anderson, J. R., and Fujita, K. (2006). Redundant food searches by capuchin monkeys (*Cebus apella*): A failure of metacognition? *Animal Cognition*, 9, 110–17.

Premack, D. and Premack, A. J. (1994). Levels of causal understanding in chimpanzees and children. *Cognition*, 50, 347–62.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1, 515–26.

Schneider, W. and Sodian, B. (1988). Metamemory—memory behavior relationships in young children: Evidence from a memory-for-location task. *Journal of Experimental Child Psychology*, 45, 209–33.

Smith, J. D. (2009). The study of animal metacognition. *Trends in Cognitive Sciences*, 13, 389–96.

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R., and Erb, L. (1995). The uncertain response in the bottlenosed dolphin (*Tursiops truncatus*). *Journal of Experimental Psychology: General*, 124, 391–408.

Smith, J. D., Shields, W. E., Schull, J., and Washburn, D. A. (1997). The uncertain response in humans and animals. *Cogntion*, 62, 75–97.

Smith, J. D., Shields, W. E., and Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26, 317–73.

Smith, J. D., Beran, M. J., Redford, J. S., and Washburn, D. A. (2006). Dissociating uncertainty states and reinforcement signals in the comparative study of metacognition. *Journal of Experimental Psychology: General*, 135, 282–97.

Sodian, B. and Schneider, W. (1990). Children's understanding of cognitive cueing: How to manipulate cues to fool a competitor. *Child Development*, 61, 697–704.

Son, L., and Kornell, N (2005). Metacognitive judgments in rhesus macaques: Explicit versus implicit mechanisms. In H. S. Terrace and J. Metcalfe (Eds.) *The missing link in cognition: origins of self-reflective consciousness*, pp. 298–320. New York: Oxford University Press.

Suda-King, C. (2008). Do orangutans (*Pongo pygmaeus*) know when they do not remember? *Animal Cognition*, 11, 21–42.

Washburn, D. A., Smith, J. D., and Shields, W. E. (2006). Rhesus monkeys (*Macaca mulatta)* immediately generalize the uncertain response. *Journal of Experimental Psychology: Animal Behavior Processes*, 32, 85–9.

Chapter 5

The emergence of metacognition: affect and uncertainty in animals

Peter Carruthers and J. Brendan Ritchie

This chapter situates the dispute over the metacognitive capacities of non-human animals in the context of wider debates about the phylogeny of metarepresentational abilities. We first clarify the nature of the dispute, before contrasting two different accounts of the evolution of metarepresentation. One is first-person-based, claiming that it emerged initially for purposes of metacognitive monitoring and control. The other is social in nature, claiming that metarepresentation evolved initially to monitor the mental states of others. These accounts make differing predictions about what we should expect to find in non-human animals: the former predicts that we should find metacognitive capacities in creatures incapable of equivalent forms of mindreading, whereas the latter predicts that we should not. We elaborate and defend the latter form of account, drawing especially on what is known about decision making and metacognition in humans. In doing so we show that so-called 'uncertainty monitoring' data from monkeys can just as well be explained in non-metarepresentational affective terms, as might be predicted by the social-evolutionary account.

Introduction: the meaning of 'metacognition'

We assume that readers of this volume will by now have some familiarity with the sorts of paradigms that have been used to provide evidence of metacognition in non-human primates. In a common type of experiment (e.g. Smith et al. 2008), animals are trained to perform a primary task such as making a discrimination of some sort between categories (e.g. sparse versus dense) to achieve a favoured reward (either immediately, or after a delay; Couchman et al. 2010). After training, the animals are also provided with an 'opt out' response of some kind, which they tend to use in difficult cases where they are more likely to make (or have made) an incorrect judgement. Opting out generally either avoids the penalty that accompanies a mistaken answer (such as a time-out before there is another opportunity to obtain a reward), or guarantees a less-favoured reward. Such results are said to show that the animals are aware of their own uncertainty, especially since similar use of the opt-out response in humans is associated with self-attributions of uncertainty.

We fully accept that this body of work, taken as a whole, cannot be explained in low-level associationist terms, as involving mere conditioned responses to stimuli. A great deal of careful experimentation has been done to demonstrate that this is not the case, and we are happy to embrace this conclusion (Beran et al. 2009; Couchman et al. 2010; Smith et al. 2010; Washburn et al. 2010). So it should be agreed that the animals have beliefs about the contingencies of the experiment and take executively-controlled decisions that depend on those beliefs (as well as having goals and other states like emotions, which some have been reluctant to attribute to animals; but see Panksepp 2005).

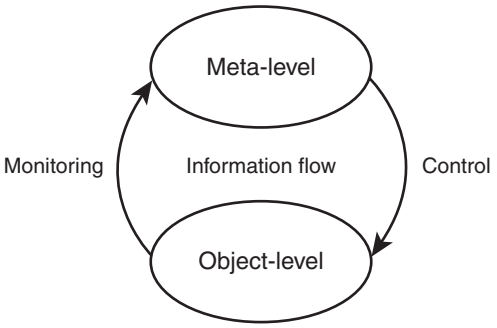


Fig. 5.1 Metacognitive monitoring and control (adapted from Nelson and Narens 1990).

However, to say that the animals' behaviour is fully cognitive and executively controlled is not yet to say that it is *meta*cognitive, in the sense in which this term is employed throughout cognitive and developmental psychology. For metacognition is generally defined as 'thinking *about* thinking' (Flavell 1979; Dunlosky and Metcalfe 2009), and therefore as involving metarepresentation. Moreover, metarepresentation in turn is understood to require a representation that represents another representation, or a mental state whose content represents, and is *about*, another mental state.

This definition of 'metacognition' accords with the standard model for classifying and characterizing metacognitive processes in humans (Nelson and Narens 1990; see Fig. 5.1), in which a metalevel monitors, represents, and controls the processes of object-level cognitive systems. Since those who study metacognitive processes in animals often cite this model with approval (e.g. Smith et al. 2003, 2006; Couchman et al. 2010), we assume that is it some version of *this* architecture, or some of its components, that the animals in question are claimed to possess when they are said to have metacognitive capacities. And it should be noted that an important aspect of the Nelson and Narens model has always been that the metalevel contains a meta*model* of the object-level, with the metalevel containing metarepresentions of processes and events at the object-level. (See, for example, Nelson and Narens, 1990, p. 126, Principle 2.)

While these definitional issues are important, we should stress that this is only because clarity is important for the progress of science. What ultimately matters, of course, is knowledge of the cognitive structures and processes that underlie the animals' behaviour, not the words we use to express that knowledge. What we will suggest is that some of the behaviour (specifically, so-called 'uncertainty-monitoring' behaviour) that has been claimed to support the presence of a metacognitive architecture can just as well be explained in non-metarepresentational affective terms.

Moreover, it should be stressed that even if some of the processes employed by animals in uncertainty-monitoring experiments might appropriately be described in terms of 'monitoring and control', it is another matter to claim that the monitoring in question is metarepresentational, or metacognitive in the standard sense. For there are multiple monitoring and control processes in human cognition that are not metarepresentational (Metcalfe 2008). Consider, for example, the use of forward models in the control of action. When motor schemata are activated and sent to the motor system to initiate an action, an efference copy of those instructions is sent to an emulator system that constructs a forward model of the expected sensory consequences of the movement (Wolpert and Kawato 1998; Wolpert and Ghahramani 2000; Jeannerod 2006).[1]

[1] This same system is also used offline, when subjects mentally rehearse potential actions for purposes of decision-making. In such cases a motor schema is activated, and although the instructions that would normally be sent to the muscles have been inhibited, the emulator system goes ahead and constructs a representation of the expected sensory consequences. This sensory representation can be 'globally

This is then received as input by a comparator mechanism that also receives reafferent sensory feedback, issuing in swift online adjustments in the action when there is a mismatch. Note that the comparator system is entirely non-metarepresentational in nature: it receives a sensory-coded representation of the intended outcome and compares this with sensory input from the actual outcome as it unfolds. When these fail to correspond, it employs an algorithm that adjusts the motor instructions to bring about a closer match. It doesn't need to represent either the motor intention or the current experiences resulting from the action as such. No metarepresentations are needed, and no one in the field of motor processing thinks that they are employed.

We will assume, then, that those who propose metacognitive explanations for the behaviour of animals in uncertainty-monitoring experiments intend this in the standard sense: they are claiming that the animals metarepresent their own states of uncertainty, and modify their behaviour as a result. We will suggest, in contrast, that the data can equally well be explained in non-metarepresentational terms. First, however, we propose to situate the issue within a wider debate about the evolutionary emergence of metarepresentational capacities.

The phylogeny of metarepresentation

Metacognition and mindreading (or 'theory of mind') are widely believed to overlap (at least) in their psychological bases and evolutionary histories. This is because both rely, fundamentally, on metarepresentation: the representation of mental states. In the case of mindreading, this involves attributing mental states to others, while in metacognition we attribute mental states to ourselves. When one attempts to explain the adaptive advantage that these capacities supplied to our ancestors, a notion of control is invoked in each case. Mindreading allows us to predict the behaviour of others in order to control our own (social) behaviour. Hence, mindreading is thought to have evolved to navigate an increasingly complex social world, engaging with multiple conspecifics in groups with complex social organization. Metacognition, in contrast, allows us to monitor and control object-level systems in our own mind, enabling us to learn and reason more flexibly.

Metarepresentation then features in both mindreading and metacognition, but in the service of rather different functions (social cognition versus cognitive control). This leads us to ask which function of metarepresentation is evolutionarily prior (as well as how this bears on the question of human cognitive architecture). The question of prioricity naturally suggests two kinds of account of the evolution of metarepresentational capacities.[2]

According to one approach, the capacity to represent one's own mental states (or some subset thereof) evolved first (Couchman et al. 2009), presumably to enable animals to accrue the benefits of metacognitive monitoring and control. Once evolved, the conceptual and inferential resources involved were later exapted for attributing mental states to other agents. There are two main ways in which this could have happened, partly motivated by different views of human mindreading. Either these first-person resources were redeployed to form the basis of a distinct mindreading faculty of the sort defended by Nichols and Stich (2003), or they were combined with emerging capacities for imaginative perspective-taking to enable *simulations* of the mental lives of others (Goldman 2006). We will refer to these as 'first-person-based' accounts of the

broadcast' (in the sense of Baars, 1988) when attended to, thus being made available to a range of systems to draw inferences and evaluate the action. We return to these points later in the chapter.

[2] We assume that no one should now think that these capacities result from general learning, and that everyone should agree that they are innately channelled in development to some significant degree. While these assumptions go undefended here, they are in fact supported by large and varied bodies of data. See Carruthers (2011) for further discussion.

evolution of metarepresentation, while making no attempt to adjudicate between dual-mechanism and simulationist variants.

According to the alternative approach, a capacity to attribute mental states to other agents evolved first, driven by the exigencies of social living and resulting in an innately channelled mindreading faculty of some sort. But this mindreading-based account also admits of two main variants. According to one, a core *capacity* to make self-attributions would have been present from the start, since there would have been nothing to prevent subjects from turning their mindreading abilities on themselves, treating the self as an agent like any other. A disposition to attribute mental states to oneself on a regular basis would only have required the motivation to direct one's attention accordingly (Carruthers 2011). According to the other variant of a mindreading-based account, in contrast, some sort of self-monitoring mechanism was subsequently added to the third-person mindreading system, enabling direct (non-sensory) access to one's own mental states (Frith and Happé (1999) seem to have in mind something like this). In this case we propose *not* to remain neutral between the two variants, but will work with a self-directed-mindreading account throughout. This provides the cleanest contrast with first-person-based approaches. And there is, in fact, a good deal of evidence against the monitoring-mechanism alternative (see Carruthers 2011).

We will shortly compare the first-person-based and mindreading-based accounts of the evolution of metarepresentation with respect to the predictions that each makes regarding the comparative data. But first it is worth noting an apparent anomaly for the former. This is that it is widely agreed among psychologists that human metacognitive capacities (or at least those of an uncontroversially metarepresentational sort) are far from impressive. For example, one robust finding in the literature is that people's metacognitive judgements of learning are only moderately correlated, at best, with later recall (Leonesio and Nelson 1990; Dunlosky and Metcalfe 2009), and another is that correlations between metacognitive judgements of text comprehension and tests of understanding are often close to zero (Lin and Zabrucky 1998; Maki and McGuire 2002). Moreover, human metacognitive capacities are fragile and cue-based, late to develop in childhood, and are heavily dependent upon individual differences in personality and local cultural mores for their effectiveness (Stanovich and West 2000; Koriat et al. 2006, 2008; Stanovich 2009).

These findings are not what might be expected if metacognitive abilities had a long evolutionary history and are innately channelled in development. In contrast, everyone agrees that human mindreading capacities are remarkably good (although admittedly we lack any shared metric for comparing mindreading capacities with metacognitive ones). More importantly, we now have ample evidence of their early emergence in human infancy (Southgate et al. 2007, 2010; Surian et al. 2007; Song et al. 2008; Buttelmann et al. 2009b; Scott and Baillargeon 2009; Scott et al. 2010). This is just as might be predicted by a mindreading-based account of the evolution of metarepresentational abilities.

It could be replied, of course, that biological structures need only deliver small adaptive advantages in order to be selected for, especially over a long time-frame. And it is possible that metarepresentational capacities evolved initially for first-person metacognitive uses, after which the main adaptive pressure became a social one. This would explain the seemingly poor metacognitive capacities of humans combined with excellent mindreading. One might expect, however, that if metacognitive capacities had been selected for among our ancestors, then they would have come under additional adaptive pressure (leading to further robustness and reliability) when learning and decision-making become increasingly complex through the evolution of the hominin line. In any case the contrast between human native capacities for metacognition, on the one hand, and mindreading, on the other, appears striking, and provides some indirect support for a mindreading-based account of the evolution of metarepresentation.

Predictions for comparative psychology

If metarepresentational capacities evolved initially for metacognitive monitoring and control, then one might expect to find creatures capable of metacognition who are *in*capable of mindreading (or at least, who are incapable of mindreading of a sort that requires equivalent metarepresentational resources; see later). At any rate, on this view there must once have been such creatures. Moreover, if creatures of this sort were now discovered, then it would provide significant support for a first-person-based account of the emergence of metarepresentation. For the mindreading-based account predicts, in contrast, that metarepresentational capacities should emerge in parallel for self and other (while perhaps allowing that other-directed metarepresentation might precede equivalent forms of metacognition, if, for example, the animals aren't initially motivated to attend to their own mental states). This is because metacognition is held to result from (or at least to employ the conceptual and computational resources of) self-directed mindreading.

The qualification about 'equivalent metarepresentational resources' is important. This is because it is widely agreed among developmental psychologists that mindreading admits of two distinct varieties, which emerge at different points in the course of infant development (Wellman 1990; Leslie 1994; Baron-Cohen 1995; Gopnik and Meltzoff 1997; Song and Baillargeon 2008). One is a form of goal/perception/knowledge–ignorance psychology that appears during the first year of life. Infants at this stage can represent the goals of other agents, as well as track what aspects of the world those agents do and do not have perceptual access to. As a result, infants at this age form appropriate expectations of agents who act in states of knowledge or ignorance respectively. But at this stage (generally referred to as 'Stage 1'), infants are incapable of representing the false belief of another agent, or of forming expectations based on how things *appear* to the other agent. These latter capacities only emerge toward the end of the fourth year of life (in language-based tasks), or by the middle of the second year of life (when non-verbal measures of competence are employed). Moreover, it is widely believed that the difference between Stage 1 and Stage 2 mindreading is one of domain-specific conceptual and/or computational *competence*, rather than resulting merely from performance factors. For it is thought that the capacity to pass Stage 2 tasks depends on an appreciation that mental representations can be *incongruent* with reality (as in a false belief), as opposed to merely *omitting* an aspect of reality (as happens with ignorance).[3]

There is now evidence of Stage 1 mindreading capacities in non-human animals, not only among other primates such as chimpanzees and rhesus macaques (Hare et al. 2000, 2001, 2006; Flombaum and Santos 2005; Melis et al. 2006; Santos et al. 2006; Buttelmann et al. 2007, 2009a), but also among canids (dogs and wolves; Hare and Tomasello 2005; Hare 2007; Udell et al. 2008), and corvids (jays, rooks, crows, and the like; Bugnyar and Heinrich 2005, 2006; Dally et al. 2006, 2009; Bugnyar et al. 2007; Stulp et al. 2009). Note that all of these animals live in complex social groups, suggesting that the pressures of social living might have converged on the evolution of simple forms of mindreading in widely separated species (Emery and Clayton 2004), consistent with a version of the 'Machiavellian intelligence' hypothesis (Byrne and Whiten 1988, 1997).

[3] It may yet turn out that this assumption is mistaken. Rather than reflecting differences in mindreading competence, the differences in performance might turn out to result from the differing executive demands of Stage 1 and Stage 2 tasks (Carruthers, forthcoming). If so, then the failures of non-human primates on Stage 2 tasks might likewise result from problems of executive function. This would mean that the metacognitive data are incapable of adjudicating in the dispute between first-person-based and mindreading-based accounts of the evolution of metarepresentation. For there would then be no reason to think that non-human primates are capable of forms of metacognition that outstrip their capacities for mindreading, even if they employ Stage 2 metarepresentational capacities in metacognitive tasks.

Given the presence of Stage 1 mindreading in non-human primates, the finding that they may be capable of monitoring their own desires (Evans and Beran 2007), their own perceptual access (Call and Carpenter 2001; Hampton et al. 2004; Krachun and Call 2009), and their own knowledge and ignorance (Hampton 2001, 2005), fails to adjudicate in our dispute. For these findings are consistent with both self-directed-mindreading and first-person-based accounts.[4]

In contrast, the current consensus among comparative researchers is that no primate species other than humans is capable of 'Stage 2' mindreading, which would include a capacity to attribute false beliefs to other agents. For all tests of such abilities have proved negative, even when conducted in competitive situations, and even when paired with knowledge–ignorance tasks that the animals pass (Hare et al. 2001; O'Connell and Dunbar 2003; Kaminski et al. 2008; Krachun et al. 2009). So if other primates can attribute such states to themselves, then this would present an anomaly for a mindreading-based account, while providing corresponding support for a first-person-based view.

While there is no data of quite this kind in the literature, a substantial body of work on uncertainty monitoring aims to show that members of many primate species are capable of monitoring their own states of certainty and uncertainty, and of choosing adaptively as a result. This might be taken to demonstrate that these animals are capable of Stage 2 metacognition, suggesting that they possess the *concept* of false belief, at least, and can apply it in the first person. For one might think that mastery of the concept of uncertainty requires a capacity to understand that one's beliefs are potentially false. Whether or not this is so will be discussed in the next section.

Uncertainty and feelings of uncertainty

Uncertainty, like certainty, is fundamentally a cognitive state, not an emotional one. To be certain of something is to have a high degree of belief that it is the case. (This might be realized in the form of an especially strong signal produced by a classifier mechanism, for example, or an especially strong memory trace.) To be uncertain of something is to have a low degree of belief that it is so (perhaps realized in a weak signal from a classifier mechanism, or a weak memory trace). However, each of these states can also give rise to distinctive emotional feelings of confidence or uncertainty. Moreover, each will have other cognitive and behavioural effects as well, including *fluent* cognitive processing (in the case of certainty) and *disfluency* (in the case of uncertainty).[5] These further consequences of uncertainty will be used to undergird our alternative (non-metarepresentational) explanations of the uncertainty-monitoring data in the next section.

If animals self-monitor and metarepresent themselves as uncertain of something, then they must be representing that they have a low degree of belief in it. This will require that they possess Stage 2 metarepresentational resources. For self-attribution of ignorance cannot be sufficient for representing that one is certain of something (utilizing one of the concepts from Stage 1), and nor can thinking that one is ignorant be sufficient for representing uncertainty. This is because neither knowledge nor ignorance admit of degrees, and nor do they imply some level of incongruency with the world, as do degrees of belief. (Recall that a capacity to represent that a mental state

[4] In fact we have doubts about the strength of some of this evidence. In particular, success in the memory monitoring experiments conducted by Hampton (2001) does not require attribution of knowledge or ignorance to oneself. It just requires the presence or absence of memory. The animal needs to act in one way if a memory is present, and to act in another if it is not. But in neither case does it need to entertain a metarepresentation of memory. See Carruthers (2008).

[5] Cognitive fluency is the ease with which information is processed in the mind, and is signalled by such factors as the speed with which a decision is reached or an item is recognized.

is incongruous with the world is thought to be the hallmark of Stage 2 metarepresentation.) Moreover, in principle the metarepresentational states involved could be based on self-monitoring that is direct (detecting or introspecting a judgement with a low degree of belief) or indirect (detecting and classifying sensory or behavioural cues of the underlying state of uncertainty). Since no one in the human metacognition literature thinks that monitoring is direct, we propose to dismiss this possibility in respect of animals also (Koriat 2000; Dunlosky and Metcalfe 2009).

In fact it should be stressed that there is general agreement among researchers that human metacognitive judgements are *cue based* (Dunlosky and Metcalfe 2009). Judgements about whether one has learned something or whether one knows something are grounded in sensorily-accessible and affective cues, such as the ease with which the item in question is processed or the feeling of familiarity induced by its presentation. For although Hart (1965) once proposed a sort of direct-access model in order to explain feelings of knowing, his account has attracted very little empirical or theoretical support since then (Koriat 2000). We should therefore expect that animals, too, will need to base their judgements on indirect cues—perhaps their own disfluency, or perhaps their own feelings of uncertainty.

Since humans in uncertainty-monitoring experiments must base their reports of their uncertainty on sensorily-accessible cues of some sort, it is reasonable to assume that the same, or something similar, is true of non-human primates. So it will be important to know how feelings of uncertainty should be characterized, as well as what other similar cues might be in the offing. What we can say with confidence is that often the feelings in question are negatively-valenced states accompanied by a degree of arousal that is proportional to what is at stake. Feeling uncertain can feel bad (to a greater or lesser degree), and it can also be agitating when concerned with important matters.[6]

Feelings of uncertainty are caused by underlying states of uncertainty (that is, low degrees of belief). It is a separate question, however, what the negative valence component of the feeling is directed toward. What is it that one feels bad about, when one feels uncertain? What situation or state of affairs is it that *seems bad* as a result of negative valence, in the way that fear makes the threatening object seem bad and anger makes the causes of damage to oneself or to one's own seem bad? One possibility would implicate metarepresentation in the very feeling of uncertainty itself, utilizing metarepresentational resources. It may be that what strikes one as bad is that one has a low degree of belief. On this account, a judgement to the effect that one has a low degree of belief would be built into (or at least accompany) the feelings in question, providing the intended object or target of those feelings.

What we propose, however, is that feelings of uncertainty (in both humans and animals) are more plausibly seen as directed at the world (in particular, at the primary options for action that are open to one), rather than at one's own mental states. Consider what happens when people engage in the Iowa Gambling Task, for example (Bechara et al. 1994). Subjects are required to select from four decks of cards with different probabilities of winning or losing. Two of the decks produce steady gains in the long-run (while sometimes issuing in big losses), while two produce long-term losses (and yet sometimes issue in big gains). After a while subjects begin to make most of their selections from the 'good' decks, but before they are capable of explicit recognition that those decks are better (let alone capable of articulating *why* they are better). Presumably, as a

[6] Note that we are not claiming that there is a unique introspectively-accessible feeling that is distinctive of states of uncertainty. Nor do we think that affective changes are always consciously experienced. All we need to be committed to for present purposes is that there will generally be *some* degree of affective change accompanying states of uncertainty, whether consciously experienced or not, and that these can exert an influence on subsequent behaviour.

result of previous learning, the good decks are unconsciously appraised as more likely to issue in gains. As a result, the thought of selecting from those decks is positively valenced, making those options seem better. But in addition, some minor degree of arousal is also present, since subjects display an increased galvanic skin response when reaching toward one of the bad decks.[7]

We should stress that in cases of this sort the affective changes can be quite minor, and may pass unnoticed by the subject. Yet still the good options seem good and the bad options seem bad, with effects on behaviour that can be quite significant. Certainly in humans, minor forms of affective priming can have large behavioural consequences. For example, Winkielman et al. (2005) used briefly presented, backward-masked, happy and angry faces (which were never consciously perceived) before subjects sampled a novel beverage. Thirsty subjects primed with positive affect drank twice as much of the beverage as those primed with negative affect, and in another condition, they offered to pay twice as much for a can of the drink having taken just a sip. Yet these unconscious primes had no discernable effects on the subjects' mood.

In fact we think that uncertainty-based decision-making may be best understood as of-a-piece with affectively-based decision-making generally, of the sort characterized by Damasio (1994), Gilbert and Wilson (2007), and many others. On this kind of account one runs the instructions for a motor action offline, using the efference copy to generate a forward model of its outcome (as described in the first section). When attended to, this is globally broadcast as an imagistic representation of the action, which one's evaluative and emotional systems receive and respond to. The result is some degree of positive or negative affect, which provides the motivation to execute the action or to seek an alternative means to the goal (or to pursue an alternative goal). On this kind of account feelings of uncertainty would consist of negatively valenced affect that is caused by the thought of an otherwise-attractive action, and that is directed toward the situation represented in the content of that thought. (It is the performance of the action that seems bad as a result, not the fact that one is thinking about it.)

There is some reason to believe that members of other primate species might be capable of such processes of mental rehearsal and affective evaluation, underlying their limited capacity for advance planning (Sanz et al. 2004; Mulcahy and Call 2006), and perhaps also explaining instances of 'insight' behaviour (see Carruthers 2006, for discussion). And indeed, a similar capacity might be more widespread still. Think of the cat that crouches down as if to leap, *literally* rehearsing (the first stages of) a difficult leap from a roof to a nearby tree. Presumably the act of representing the action issues in appraisals of likely success, resulting either in positive affect (felt confidence) directed at the intended leap, or in negative affect (felt uncertainty), leading the cat to seek other solutions.

In this section we have distinguished uncertainty from the cognitive and affective consequences of uncertainty, and we have pointed out that animals, like humans, will need to rely on indirect cues of uncertainty, even if they do metarepresent such states. We have also suggested that the valence component of feelings of uncertainty is directed at the primary response options, rather than at one's own mental states. While humans engage in many forms of metacognitive

[7] Amiez et al. (2003) used a decision-making task equivalent to the Iowa Gambling Task with macaques, but found that the galvanic skin response occurred *after* the animals had made their selection, seemingly in anticipation of a reward. Quite how galvanic skin responses in uncertainty tasks like these are supposed to support Bechara and colleagues' own 'somatic marker' account of affective decision-making is a complicated matter, however (Dunn et al. 2006). So it is far from clear that this result undermines their hypothesis. But in any case our view is not committed to the details of this particular theory of the manner in which affective cues influence decision-making. Indeed, our primary focus is on the valence component of affect, rather than on bodily arousal.

decision-making, requiring them to metarepresent their own mental states and processes, basic forms of affectively-based decision-making are *not* metarepresentational in humans. When we represent and respond affectively to alternative courses of action, no metarepresentations need be involved. As a result, in the following section we will suggest that the uncertainty-monitoring data may be explained without ascribing metarepresentational capacities of any sort to the animals involved.[8]

Affective explanations of the evidence

The present section will discuss three distinct non-metarepresentational explanations of the uncertainty-monitoring data from non-human primates. The first is unsatisfying on its own. But each of the others provides a viable alternative to a metarepresentational account. We will focus especially on a valence-based theory that builds on some of the ideas from the previous section.

Degrees of belief

One form of non-metarepresentational explanation is proposed by Carruthers (2008), who appeals to degrees of belief and desire, together with ordinary practical reasoning, to show how the uncertainty-monitoring data can be explained. While this account may not be incorrect, it strikes us as incomplete. This is because it is purely cognitive in nature, and fails to provide for the emotional character of uncertainty.[9] Since humans in such experiments report not only that they *are* uncertain (in the sense of having low degrees of belief) but that they *feel* uncertain (and indeed, since a judgement that one is uncertain must be grounded in indirect cues such as feelings of uncertainty), it seems inadvisable to omit an affective component from the explanation. For the results of uncertainty monitoring experiments with humans can parallel the animal uncertainty-monitoring data quite closely (Smith et al. 2003, 2008; Smith 2005). Accordingly, two further accounts will be outlined here. Each appeals to the consequences of states of uncertainty, while differing from one another in the factors that are utilized. We should emphasize, however, that these accounts are consistent with one another. Each might apply in different kinds of case, or they might combine together in the same cases.

Affective consequences as cues

One possibility is that the animals in question have learned to use some aspect of their own feelings of uncertainty as a *cue*, but without at any time metarepresenting that they are uncertain (i.e. without categorizing their affective experience as a feeling *of* uncertainty), or thinking that their judgements or memories are likely to be false. In effect, they may be a following a rule like, 'When in a state of *that* sort [uncertainty], opt out and do something different', which would only require possession of an indexical, non-mental, concept. This can explain why the animals are more likely to press the opt-out key in psychophysically difficult cases, and it can also explain how the animals are able to generalize the use of the opt-out key when presented with it in the context of a newly learned discrimination task (Son and Kornell 2005; Kornell et al. 2007). But neither the feeling itself, nor the indexical concept used to identify it, need involve metarepresentation.

[8] This will mean that even if the evidence suggesting that non-human primates are capable of Stage 1 forms of mindreading proves to be unsound, it will still be the case that the uncertainty-monitoring data fail to support first-person-based views. For uncertainty-monitoring behaviour arguably fails to involve metarepresentations of any sort (whether Stage 1 *or* Stage 2).

[9] Emotions might, however, be incorporated into the account as a way of implementing the so-called 'gating mechanism' appealed to.

Rather, just as humans are apt to do (Koriat 2000), the animals might utilize their own disfluency or the bodily feelings associated with uncertainty as cues when confronted with difficult cases.

We know from the human case that affective states provide experiential cues for metacognitive deliberation (Koriat 2000; Koriat et al. 2006, 2008). For example, differences in affective states during learning or retrieving information are used as cues that reflect the underlying processing dynamics or processing fluency. In the case of information that is easy to process, greater fluency results, causing positive affect (Winkielman and Cacioppo 2001). The extent to which people show confidence in automatic, intuitive, judgements is heavily dependent on processing fluency. This is true, for example, in low-difficulty recognition tasks, where selection can be based primarily on the feeling of familiarity without serial recall (e.g. Mandler 1980), and also in implicit learning tasks (Gordon and Holyoak 1983).[10]

Disfluent processing, by contrast, has been suggested to play a role in initiating a transition from more automatic processing to more executively-controlled explicit processing (Alter et al. 2007). This role for disfluency has been interpreted as a cue for metacognitive processing, but being sensitive to disfluency need not presuppose any capacity for metarepresentation. For what the experiment by Alter and colleagues shows is that disfluent processing causes changes in *attention*, issuing in different forms of cognitive control. And these changes in attention to the task might drive the selection of different reasoning strategies in the absence of metacognitive processing. Alternatively, disfluent processing might cause subjects to attend to their own increased arousal, for example, which is taken as a cue to reason differently.[11]

Note that although this sort of account need not involve metarepresentation, it does rely on self-directed forms of attention. For the animals will attend to, and notice, something about themselves (such as their own bodily feelings) in order to learn the cue-based rule in question. So it can appropriately be described as a form of uncertainty *monitoring*, even if the monitoring involved is not metarepresentational. The remaining form of affect-based explanation, in contrast, is entirely outward-looking or world-directed in character, while likewise finding a basis in what is known about human decision-making.

Directed valence

Suppose that animals, like humans, integrate probabilistic information with intended goal outcomes to issue in appraisals of the likelihood of success of the options available to them in a decision-making context. In that case, when an animal has a low degree of belief in something (that the pattern on the screen is dense rather than sparse, or that it has just touched the longest of the lines on the screen, for example), then actions that depend upon the truth of that belief will be appraised as unlikely to succeed.[12] Consequently the animal will experience some degree of anxiety when it contemplates pressing the 'dense' key or the 'gamble' option (albeit quite minor, since the stakes are so low). With negative valence directed at the action in question, it will to that extent seem bad or aversive. In such circumstances the primary response options will be seen in a mixed evaluative light. On the one hand they will seem good, since they hold out the possibility of a significant reward; but on the other hand they will seem bad, since they are appraised as

[10] Here we presuppose a processing-fluency view of feelings of familiarity, in the manner of Jacoby (1991).

[11] Note that neither interoception nor proprioception, of the sort that might underlie awareness of arousal, are metarepresentational forms of awareness (although in a loose sense they can be described as 'introspective'). Rather, they issue in awareness of properties of the body.

[12] See Balci et al. (2009) for evidence that mice, too, are capable of making swift and accurate assessments of risk. See also Gallistel et al. (2001) for evidence that rats are excellent at tracking random changes in the probability of reward.

unlikely to succeed. The opt-out response, in contrast, will be seen as an unopposed weak positive, since it either advances the animal to a new trial without a time-out or issues in a guaranteed less favoured reward. It is not surprising, then, that the animals should press the opt-out key more often in such circumstances.

As we noted earlier, this explanation coheres well with what is known about the decision-making processes employed by humans. When humans are confronted with choices they will generally rehearse the actions involved in implementing those choices. These representations, when taken as input by the individual's affective mechanisms, will result in some degree of positive or negative affect directed at the option in question. This makes that option seem either good or bad (attractive or aversive), in many cases issuing in a decision (unless the subject opts to engage in more explicit reflection of some sort).[13]

Smith (2005) makes much of the fact that humans in uncertainty-monitoring experiments have response profiles that closely parallel those of the animals (see also Smith et al. 2008). Since the humans report that they opt out in conditions of uncertainty because they are aware of being uncertain, this is said to give us reason to attribute similar awareness to the animals. But it does not. For basic forms of decision-making in humans don't employ metarepresentational awareness, as we have seen. So both humans and animals will experience negatively valenced forms of anxiety directed toward the primary response options, resulting from an appraisal of low likelihood of success. (The latter in turn is grounded in the low degree of belief that attaches to the categorization or judgement underlying the required discrimination.) This will make those options appear bad or mildly aversive. Such perceptions, when strong enough, will leave the opt-out option as the better-seeming alternative. All of this is entirely non-metarepresentational, as we have noted. But humans, with their highly-developed mindreading capacities, will categorize the state they are in *as* a feeling of uncertainty, either automatically or when asked to explain their choice. This categorization might play no role, however, in their basic decision-making behaviour (unless it is first articulated and treated as a commitment). Indeed, their metacognizing might be largely post hoc.

What we suggest, then, is that in humans both uncertainty and its influence on behaviour should be dissociable from metarepresentational awareness of uncertainty. To the best of our knowledge this has not been directly tested. But we predict that subjects who have difficulties with mindreading (including those suffering from autism or schizophrenia) might show capacities to make adaptive use of the opt-out key in uncertainty-responding experiments that are spared in comparison with their capacity to identify themselves *as* uncertain. For example, in one condition subjects might perform the task without making any explicit metacognitive judgements, whereas in another they might be required to make such a judgement before deciding whether or not to opt out. Our prediction is that performance in the former condition should be significantly better than performance in the latter, in these populations.

Further consequences of the accounts

Notice that both of the affect-based explanations mooted here make significant executive demands on the animals in question. In order for feelings of uncertainty to be used as cues to opt out, they have to be attended to. And in order for one to feel anxious at the thought of taking a particular action, that action has to be mentally rehearsed. We should predict, then, that the animals are

[13] Note that this account deviates slightly from that provided by Damasio (1994), who places more emphasis on the arousal and other bodily components of affect, rather than on the valence component as we do here. For discussion and defence, see Carruthers (2011). And note, too, that even if arousal *is* involved it can be quite subtle, perhaps depending on what Damasio calls '*as if* affect.'

unlikely to make adaptive use of opt-out behaviour in cases where they are required to execute some concurrent task. Note that this prediction is not made by the degrees-of-belief account alone (independent of any role for epistemic emotions). However, it is also a prediction of the metarepresentational account. So the finding that use of the opt-out response diminishes when animals are required to engage in an ancillary task (Smith 2011), does nothing to support a metarepresentational account of uncertainty monitoring over its affect-based competitors.

Moreover, each of the affect-based accounts makes the following empirical prediction. Mood manipulations that are effective in reducing anxiety, or drugs that produce such an outcome, should significantly reduce the extent to which animals opt out in conditions of uncertainty. In contrast, the degrees-of-belief account fails to make any such prediction. For it is purely cognitive in nature. Moreover, any metacognitive account that is cast in purely cognitive terms (merely maintaining that the animals are aware of their uncertainty, for example) will likewise fail to make such a prediction. However, mood manipulations, even if successful, would not necessarily support an affect-based account of uncertainty-monitoring behaviour over a metarepresentational one. For metacognitive theorists can presumably claim that what is represented is an *emotional* state of uncertainty, and in that case manipulations that reduce anxiety will have the effect of making it harder to monitor and metarepresent the relevant state.

None of the tests that have been employed to date are capable of discriminating between metarepresentational and non-metarepresentational explanations of uncertainty-monitoring behaviour. So we are forced to fall back on indirect reasons that might favour one or other kind of explanation, of the sort that have been in play up to now. Some further considerations of this kind will form the topic of the next section.

Species differences and individual variation

Smith (2005) and others have argued that differences in uncertainty-monitoring behaviour between species support a metarepresentational account. In this section we challenge this interpretation, while also arguing that individual differences in such behaviour among humans may be problematic for first-person-based accounts to accommodate.

Species differences

We agree that differences in uncertainty monitoring behaviour across species favour a metarepresentational account over an associative learning competitor, since the species that fail in these tasks (rats and pigeons) excel at such learning (Smith 2005; Smith et al. 2009). But they don't support a metarepresentational account over either of the affectively-based proposals discussed earlier in the 'Affective explanations of the evidence' section. This is because there may be species differences in the extent to which anxiety is created in foraging situations, or differences in the extent to which members of a given species pay attention to or notice their own bodily feelings, or differences in capacities to engage in mental rehearsal of action. None of these differences is yet confirmed. But until they are ruled out, we have no positive reason to believe that the difference between the species is a metarepresentational one.

It might be claimed that differences in uncertainty-monitoring behaviour among distinct species of monkey provide a greater challenge for non-metarepresentational accounts (Beran et al. 2009; cf. Basile et al. 2009). Capuchin monkeys, in particular, rarely if ever make use of the opt-out response, even after numerous trials, and even under conditions designed to bias the monkeys toward using the opt-out response. Macaque monkeys, in contrast, show response profiles that closely parallel those of humans. It should be obvious from the previous discussion, however, that there are multiple types of resource that could potentially be used to explain these differences

without appealing to metarepresentational capacities, and some of these explanations are independently plausible.[14]

It is possible that the two species differ in the extent to which they are apt to experience anxiety in foraging situations. In particular, if capuchins feel little or no anxiety when confronted with a difficult discrimination task to gain a food reward, then they will not be motivated to use the opt-out key. If macaques are more like humans in this respect, however, then the primary response options will be experienced as aversive in cases of difficult discrimination, making it more likely that the animals will use the opt-out response. Alternatively, capuchins might experience anxiety, but not know what to do with it (i.e. what control operation to adopt). (Compare people who are used to dealing with high degrees of anxiety and those who are not.) These suggestions could be motivated by the following ecological facts.

Capuchins are arboreal, living locally in forest environments that provide ample sources of fruits, nuts, leaves, and insects that constitute their primary diet. Although they experience food competition within groups, adults are known to share food with unrelated infants, and adults will often share food with one another (De Waal 2000). Macaques, in contrast, are often semi-nomadic with broad ranges, and have colonized a wide set of ecologies, with the largest distribution of any non-human primate genus (Fleagle 1998). Illustrating their flexibility in adapting to new environments, 'weed' macaques (such as the rhesus macaque) have been able to thrive in human environments (Richard et al. 1989). Although they, too, are omnivorous, they are subject to intense food competition within groups (Sterck and Steenbeek 1997). It would not be surprising, then, that they might have become adapted to experience and deal with anxiety in difficult foraging situations, since they face far more uncertainties when foraging than do capuchins.

Individual differences

Smith (2005) also notes that both humans and the other primates in these experiments display similar ranges of individual difference. Some people, and some animals, never make use of the opt-out key, and confine themselves to the primary response options, whereas others opt out adaptively in circumstances where they are likely to make (or to have made) a mistake. It is unclear why this should be thought to support a metarepresentational account, however. (Indeed, we will suggest in a moment that it may cause problems for that account.) In any case, each of the two affect-based theories is capable of explaining this fact.

In the first place, it is well known that there are chronic differences among people in the extent to which they pay attention to the bodily (arousal) component of their emotional states (Barrett 1998; Gasper and Clore 2000; Barrett et al. 2004), and one might expect the same to be true of other primates. Such individuals are unlikely to notice the manifestations of their own state of uncertainty, and so will be less likely to learn to use them as cues to opt out. It is also well known that there are chronic differences between people (and presumably other primates) in the extent to which they become anxious in everyday situations. Those who aren't easily made anxious will fail to see the primary response options as bad or aversive, and so will lack any motivation to use the opt-out response, whereas those who are more easily made anxious will opt out more often.

[14] We note that Beran et al. (2009) themselves offer accounts of the failure of capuchins in these tasks that don't seem to depend on an absence of metarepresentational capacities. They suggest, for example, that capuchins may lack the ability to appreciate the abstract and indirect benefit of selecting the uncertainty response to maximize reward, leading them to focus on the primary, directly rewarding, options. If this is transposed into a positive account of macaques' success in these tasks, then the account is no longer a metacognitive one. For appreciating an indirect benefit need not require metarepresentation.

In contrast, while all views might predict that there will be individual differences in the *extent* to which people make use of the opt-out option, the fact that some people (and animals) almost *never* employ it is harder for a metacognitive theorist to accommodate. *How* hard it is will be a function of the proportion of subjects who never opt out. If such people are rare, then they might be considered the tail-ends of a normal distribution. But if they are numerous, and the normal distribution curve is not very steep, then this will be more problematic. For recall that the conceptual and inferential resources necessary to monitor one's own mental states are claimed to have been selected for precisely because of the adaptive advantages that they yield in situations like this. It would therefore be puzzling if there should turn out to be many individuals who nevertheless fail to make use of those resources. (It would much as if we found a significant proportion of people who never make use of episodic memory.) For there is surely just as much need for people to monitor uncertainty as there ever was in our evolutionary past. If metarepresentional resources evolved, in part, to enable animals to monitor their own uncertainty and respond adaptively, then one would expect that those resources would be regularly and reliably employed by the vast majority of normal individuals. The affect-based accounts, in contrast, can appeal to widespread individual differences that are already known to exist.

It seems, then, that in the absence of a direct experimental test, there are no indirect reasons to favour a metarepresentational account of the uncertainty-monitoring data over its affect-based competitors; indeed, there are some reasons to prefer the latter.[15]

Conclusion

We conclude that existing uncertainty-monitoring experiments with non-human primates fail to discriminate between a metacognitive (metarepresentational) account and those that rely on non-metarepresentational uses of feelings of uncertainty. Until experiments that might tease apart these differing explanations have been done, a metarepresentational account of the uncertainty-monitoring data is unsupported. As a result, while we have good reason to think that these animals are capable of taking executively controlled decisions in many ways like our own, we presently have no reason to prefer a first-person-based account of the evolutionary emergence of metacognition over its mindreading-based competitor.

Acknowledgements

We are grateful to Michael J. Beran and an anonymous reviewer for their insightful comments on an earlier draft of this chapter.

References

Alter, A., Oppenheimer, D., Epley, N., and Eyre, R. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136, 569–76.

Amiez, C., Procyk, E., Honoré, J., Sequeira, H., and Joseph, J.- P. (2003). Reward anticipation, cognition, and electrodermal activity in the conditioned monkey. *Experimental Brain Research*, 149, 267–75.

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

[15] There is one final possibility that has not been discussed here. This is that epistemic emotions like uncertainty are *non-conceptual* forms of first-person metarepresentation in virtue of their function of informing us of the underlying risk of epistemic failure. See Proust (2009a, 2009b) for defence of a view of this sort. (Note, however, that Proust herself declines to use the language of 'metarepresentation' in this connection.) For a critique of this idea see Carruthers (2011).

Balci, F., Freestone, D., and Gallistel, C. R. (2009). Risk assessment in man and mouse. *Proceedings of the National Academy of Sciences*, 106, 2459–63.

Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, MA: MIT Press.

Barrett, L. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion*, 12, 579–99.

Barrett, L., Quigley, K., Bliss-Moreau, E., and Aronson, K. (2004). Interoceptive sensitivity and self-reports of emotional experience. *Journal of Personality and Social Psychology*, 87, 684–97.

Basile, B., Hampton, R., Suomi, S., and Murray, E. (2009). An assessment of memory awareness in tufted capuchin monkeys (*Cebus apella*). *Animal Cognition*, 12, 169–80.

Bechara, A., Damasio, A., Damasio, H., and Anderson, S. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50, 7–15.

Beran, M., Smith, J., Coutinho, M., Couchman, J., and Boomer, J. (2009). The psychological organization of 'uncertainty' responses and 'middle' responses: A dissociation in capuchin monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 371–81.

Bugnyar, T. and Heinrich, B. (2005). Food-storing ravens differentiate between knowledgeable and ignorant competitors. *Proceedings of the Royal Society of London B*, 272, 1641–6.

Bugnyar, T. and Heinrich, B. (2006). Pilfering ravens, *Corvus corax*, adjust their behavior to social context and identity of competitors. *Animal Cognition*, 9, 369–76.

Bugnyar, T., Stöwe, M., and Heinrich, B. (2007). The ontogeny of caching in ravens. *Animal Behavior*, 74, 757–67.

Buttelmann, D., Carpenter, M., Call, J., and Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental Science*, 10, F31–38.

Buttelmann, D., Call, J., and Tomasello, M. (2009a). Do great apes use emotional expressions to infer desires? *Developmental Science*, 12, 688–98.

Buttelmann, D., Carpenter, M., and Tomasello, M. (2009b). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112, 337–42.

Byrne, R. and Whiten, A. (Eds.) (1988). *Machiavellian Intelligence*. Oxford: Oxford University Press.

Byrne, R. and Whiten, A. (Eds.) (1997). *Machiavellian Intelligence II*. Cambridge: Cambridge University Press.

Call, J. and Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition*, 4, 207–20.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind and Language*, 23, 58–89.

Carruthers, P. (2011). *The Opacity of Mind*. Oxford: Oxford University Press.

Carruthers, P. (forthcoming). Mindreading in infancy.

Couchman, J., Coutinho, M., Beran, M., and Smith, D. (2009). Metacognition is prior. *Behavioral and Brain Sciences*, 32, 142.

Couchman, J., Coutinho, M., Beran, M., and Smith, J. D. (2010). Beyond stimulus cues and reinforcement signals: A new approach to animal metacognition. *Journal of Comparative Psychology*, 124, 356–68.

Dally, J., Emery, N., and Clayton, N. (2006). Food-caching western scrub-jays keep track of who was watching when. *Science*, 312, 1662–5.

Dally, J., Emery, N., and Clayton, N. (2009). Avian theory of mind and counter espionage by food-caching western scrub-jays. *European Journal of Developmental Psychology*, 7, 17–37.

Damasio, A. (1994). *Descartes' Error*. London: Papermac.

De Waal, F. (2000). Attitudinal reciprocity in food sharing among brown capuchin monkeys. *Animal Behaviour*, 60, 253–61.

Dunlosky, J. and Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage Publications.

Dunn, B., Dalgleish, T., and Lawrence, A. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience and Biobehavioral Reviews*, 30, 239–71.

Emery, N. and Clayton, N. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science*, 306, 1903–7.

Evans, T. and Beran, M. (2007). Chimpanzees use self-distraction to cope with impulsivity. *Biology Letters*, 3, 599–602.

Flavell, J. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34, 906–11.

Fleagle, J. (1998). *Primate Adaptation and Evolution*, 2nd Edn. New York: Academic Press.

Flombaum, J. and Santos, L. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology*, 15, 447–52.

Frith, U. and Happé, F. (1999). Theory of mind and self-consciousness: what is it like to be autistic? *Mind and Language*, 14, 1–22.

Gallistel, R., Mark, T., King, A., and Lantham, P. (2001). The rat approximates an ideal detector of rates of reward. *Journal of Experimental Psychology: Animal Behavior Processes*, 27, 354–72.

Gasper, K. and Clore, G. (2000). Do you have to pay attention to your feelings to be influenced by them? *Personality and Social Psychology Bulletin*, 26, 698–711.

Gilbert, D. and Wilson, T. (2007). Prospection: Experiencing the future. *Science*, 317, 1351–4.

Goldman, A. (2006). *Simulating Minds*. New York: Oxford University Press.

Gopnik, A. and Meltzoff, A. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gordon, P. and Holyoak, K. (1983). Implicit learning and generalization of the 'mere exposure' effect. *Journal of Personality and Social Psychology*, 45, 492–500.

Hampton, R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of The United States of America*, 98, 5359–62.

Hampton, R. (2005). Can Rhesus monkeys discriminate between remembering and forgetting? In H. Terrace and J. Metcalfe (Eds.) *The Missing Link in Cognition*, pp. 272–95. Oxford: Oxford University Press.

Hampton, R., Zivin, A., and Murray, E. (2004). Rhesus monkeys (*Macaca mulatta*) discriminate between knowing and not knowing and collect information as needed before acting. *Animal Cognition*, 7, 239–46.

Hare, B. (2007). From nonhuman to human mind: What changed and why? *Current Directions in Psychological Science*, 16, 60–4.

Hare, B. and Tomasello, M. (2005). Human-like social skills in dogs? *Trends in Cognitive Sciences*, 9, 439–44.

Hare, B., Call, J., Agnetta, B., and Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behavior*, 59, 771–85.

Hare, B., Call, J., and Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behavior*, 61, 139–51.

Hare, B., Call, J., and Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, 101, 495–514.

Hart, J. (1965). Memory and the feeling of knowing experience. *Journal of Educational Psychology*, 56, 208–16.

Jacoby, L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–41.

Jeannerod, M. (2006). *Motor Cognition*. Oxford: Oxford University Press.

Kaminski, J., Call, J., and Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, 109, 224–34.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149–71.

Koriat, A., Ma'ayan, H., and Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, 135, 36–69.

Koriat, A., Nussinson, R., Bless, H., and Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky and J. Bjork (Eds.) *Handbook of Metamemory and Memory*, pp. 117–34. Mahwah, NJ: Erlbaum.

Kornell, N., Son, L., and Terrace, H. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18, 64–71.

Krachun, C. and Call, J. (2009). Chimpanzees (*Pan troglodytes*) know what can be seen from where. *Animal Cognition*, 12, 317–31.

Krachun, C., Carpenter, M., Call, J., and Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science*, 12, 521–35.

Leonesio, R. and Nelson, T. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 464–70.

Leslie, A. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. Hirchfeld and S. Gelman (Eds.) *Mapping the Mind*, pp. 39–67. Cambridge: Cambridge University Press.

Lin, L. and Zabrucky, K. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, 23, 345–91.

Maki, R. and McGuire, M. (2002). Metacognition for text: Findings and implications for education. In T. Perfect and B. Schwartz (Eds.) *Applied Metacognition*, pp. 39–67. Cambridge: Cambridge University Press.

Mandler, G. (1980). Recognizing: The judgment of previous occurrences. *Psychological Review*, 87, 252–71.

Melis, A., Call, J., and Tomasello, M. (2006). Chimpanzees (*Pan troglodytes*) conceal visual and auditory information from others. *Journal of Comparative Psychology*, 120, 154–62.

Metcalfe, J. (2008). Evolution of metacognition. In Dunlosky, J. and Bjork, J. (eds.) *Handbook of Metamemory and Memory*, pp. 29–46. New York: Psychology Press.

Mulcahy, N. and Call, J. (2006). Apes save tools for future use. *Science*, 312, 1038–40.

Nelson, T. and Narens, L. (1990). Metamemory: a theoretical framework and new findings. In G. Bower (Ed.) *The Psychology of Learning and Information* (Vol. 26), pp. 125–73. London: Academic Press.

Nichols, S. and Stich, S. (2003). *Mindreading*. New York: Oxford University Press.

O'Connell, S. and Dunbar, R. (2003). A test for comprehension of false belief in chimpanzees. *Evolution and Cognition*, 9, 131–40.

Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition,* 14, 30–80.

Proust, J. (2009a). The representational basis of brute metacognition: A proposal. In R. Lurz (Ed.) *The Philosophy of Animal Minds*, pp. 165–83. Cambridge: Cambridge University Press.

Proust, J. (2009b). Overlooking metacognitive experience. *Behavioral and Brain Sciences*, 32, 158–9.

Richard, A., Goldstein, S., and Dewar, R. (1989). Weed macaques: The evolutionary implications of macaque feeding ecology. *International Journal of Primatology*, 10, 569–94.

Santos, L., Nissen, A., and Ferrugia, J. (2006). Rhesus monkeys (*Macaca mulatta*) know what others can and cannot hear. *Animal Behavior*, 71, 1175–81.

Sanz, C., Morgan, D., and Gulick, S. (2004). New insights into chimpanzees, tools, and termites from the Congo basin. *American Naturalist*, 164, 567–81.

Scott, R. and Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development*, 80, 1172–96.

Scott, R., Baillargeon, R., Song, H., and Leslie, A. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology*, 61, 366–95.

Smith, J. D. (2005). Studies of uncertainty monitoring and metacognition in animals and humans. In H. Terrace and J. Metcalfe (Eds.) *The Missing Link in Cognition*. Oxford: Oxford University Press.

Smith, J. D. (2011). Presentation at the American Association for the Advancement of Science, Washington DC, 20 February 2011.

Smith, J. D., Beran, M., Redford, J., and Washburn, D. (2006). Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General*, 135, 282–97.

Smith, J. D., Beran, M., Couchman, J., and Coutinho, M. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review*, 15, 679–91.

Smith, J. D., Beran, M., Couchman, J., Coutinho, M., and Boomer, J. (2009). The curious incident of the capuchins. *Comparative Cognition & Behavior Reviews*, 4, 61–4.

Smith, J. D., Redford, J., Beran, M., and Washburn, D. (2010). Rhesus monkeys (*Macaca mulatta*) adaptively monitor uncertainty while multi-tasking. *Animal Cognition*, 13, 93–101.

Smith, J. D., Shields, W., and Washburn, D. (2003). The comparative psychology of uncertainty monitoring and meta-cognition. *Behavioral and Brain Sciences*, 26, 317–73.

Son, L. and Kornell, N. (2005). Meta-confidence judgments in rhesus macaques: Explicit versus implicit mechanisms. In H. Terrace and J. Metcalfe (Eds.) *The Missing Link in Cognition*, pp. 296–320. Oxford: Oxford University Press.

Song, H. and Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology*, 44, 1789–95.

Song, H., Onishi, K., Baillargeon, R., and Fisher, C. (2008). Can an actor's false belief be corrected by an appropriate communication? Psychological reasoning in 18.5-month-old infants. *Cognition*, 109, 295–315.

Southgate, V., Senju, A., and Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18, 587–92.

Southgate, V., Chevallier, C., and Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, 13, 907–12.

Stanovich, K. (2009). *What Intelligence Tests Miss: The Psychology of Rational Thought*. New Haven, CT: Yale University Press.

Stanovich, K. and West, R. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645–726.

Sterck, E. and Steenbeek, R. (1997). Female dominance relationships and food competition in the Sympatric Thomas Langur and long-tailed Macaque. *Behavior*, 134, 749–74.

Stulp, G., Emery, N., Verhulst, S., and Clayton, N. (2009). Western scrub-jays conceal auditory information when competitors can hear but cannot see. *Biology Letters*, 5, 583–5.

Surian, L., Caldi, S., and Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18, 580–6.

Udell, M., Dorey, N., and Wynne, C. (2008). Wolves outperform dogs in following human social cues. *Animal Behavior*, 76, 1767–73.

Washburn, D., Gulledge, J., Beran, M., and Smith, J. D. (2010). With his memory magnetically erased, a monkey knows he is uncertain. *Biology Letters*, 6, 160–2.

Wellman, H. (1990). *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Winkielman, P. and Cacioppo, J. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation leads to positive affect. *Journal of Personality and Social Psychology*, 81, 989–1000.

Winkielman, P., Berridge, K., and Wilbarger, J. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality and Social Psychology Bulletin*, 31, 121–35.

Wolpert, D. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, 3, 1212–17.

Wolpert, D. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11, 1317–29.

Chapter 6

MiniMeta: in search of minimal criteria for metacognition

Josef Perner

This is a transitory chapter linking Section I on comparative animal studies with the upcoming Section II on developmental studies with human infants and children. I am surveying the animal literature for useful methods to use with children when trying to avoid reliance on heavy metacognitive verbalization of mental states.

A sizeable research effort has evolved over the last 15 years with the aim of demonstrating metacognition in animals. The techniques involved are obviously interesting to the developmental psychologist for use on young (pre- or minimally verbal) children. Unfortunately, there are still lingering doubts (e.g. Carruthers 2008; Metcalfe 2008) as to whether these studies do show metacognition. These doubts are due to novel alternative interpretations of existing studies that have not yet been ruled out and to questions of definition. Not that there are no clear definitions, but the feeling is that these may be too restrictive for research with nonverbal organisms cutting out interesting lower level metacognitive abilities (Proust 2007, 2010; Smith 2011). My prime objective is not to arbiter between animals being or not being capable of metacognition. I want to look at the techniques to see whether they

Fig. 6.1 Minimally meta.

Haiku about meta-	**Meta-haiku**
think about thinking	A Western haiku?
or the soul's many ways	5-7-5 syllables
metacognition	that is the count

can, in principle, provide evidence of metacognition and thus have potential for adoption to use with young children. For this it is necessary in the first section to provide clarification on the different meanings of 'metacognition' and which one captures best its intuitive meaning. In the second section I elaborate on two pernicious problems for getting conclusive evidence for metacognition from behavioural investigations and illustrate these problems on some existing studies. In the third section a more systematic effort is made to look at the most promising methods and see how they fare in view of these interpretational problems. In the fourth section I explore the grey area between object-level cognition and metalevel cognition for a coherent way of identifying minimally metacognitive abilities: the MiniMeta project.

Varieties of 'metacognition'

Even when we are after minimal criteria for metacognition, we still should have some idea before the search of what it is, for which we seek criteria. We need to clarify what we mean by 'cognition', 'meta-', and their combination. To start with *cognition*: it has traditionally been used to denote one of three kinds of mental processes that deal with how the world is as opposed to those concerned with how we want the world to be and how we feel about it (Hilgard 1980: cognition—conation—affect). In Cognitive Science the term cognition has taken on a wider meaning (Fodor 1978; Wimmer and Perner 1979). Cognitive analysis of the mind assumes that all mental processes consist of transformations of representations. So not only what we know and think but also how we want the world to be and how we feel about it are open to cognitive analysis.

The term 'meta-' is Greek and means *beyond* or *after*.[1] In this sense we would interpret 'metacognition' as *special cognition*, i.e. something that goes beyond standard cognition. This very basic meaning of 'going beyond' has, however, typically been applied in a recursive fashion. For instance, *metalanguage* goes beyond language in the specific sense of *language about language*, or *metamathematics* is understood as the *study of mathematics using mathematical methods* (Wikipedia). In this tradition we end up with the meaning of 'metacognition' as *cognition about cognition*. Taking thinking as the most typical cognitive state it is also often defined as *thinking about thinking* (e.g. Flavell 1979; Dunlosky and Metcalfe 2009).

If we take cognitions the Cognitive Science way as based on representations, then, for example, a feeling about knowledge (I feel bad about not knowing the answer), would also be a metacognition.[2] Moreover, in this view metacognition implies metarepresentation, a term being worked hard in this volume. Equating metacognition with metarepresentation is considered by some an appropriate (Carruthers and Ritchie Chapter 5) but by others as too demanding a definition (Couchman et al. Chapter 1; Proust Chapter 14).

The clearest definition of metarepresentation is still the original one by Pylyshyn (1978) as *representation of the representational relationship itself*. It is however, very demanding. A representational relation relates the thing that represents (the representational vehicle—in case of a picture, the marks on the paper) with the represented, which consists of the representational target (the object or scene shown in the picture) and the representational content (the way in which the

[1] The best known word formed with meta- is probably metaphysics: the word 'metaphysics' is derived from a collective title of the 14 books by Aristotle that we currently think of as making up 'Aristotle's Metaphysics' Aristotle himself did not know the word. . . At least one hundred years after Aristotle's death, an editor of his works (in all probability, Andronicus of Rhodes) entitled those 14 books 'Ta meta ta phusika'—'the after the physicals' or 'the ones after the physical ones' (van Inwagen, 2010, p. 2).

[2] Being a clear case of metacognition hinges on the use of the word 'about'. The state we refer to with, e.g. 'feeling good about knowing the answer' is metacognition, while 'feeling good because one knew the answer,' is a debatable case. Thanks to an anonymous reviewer for alerting me to this.

picture shows the target to be: a horse in the distance as a dot; Goodman, 1976). So representing the representational relationship would at least require some sensitivity to all the relata[3]—at least that was my interpretation (Perner 1991).[4] This strict version of metarepresentation admits to true recursion, e.g. I am thinking that you are thinking that I am thinking about something. It can, therefore, also be referred to as 'recursive cognition'.

One common weaker understanding of metarepresentation is *representations that represent the content of a representation* (Leslie and Roth 1993, p. 91, M-representation; Sperber 2000, p. 117). In that case if one represents anything that does not exist (e.g. a unicorn, or a pretend scenario) then one cannot represent anything existing (no target) only the content of someone's representation. This can be the content of one's own previous thought. Any train of thought about a stable non-existing entity would then qualify as metarepresentation. In this sense metarepresentations are not truly recursive. When I think of a unicorn no recursion is possible, since the unicorn itself is not a representation (vehicle) with representational content.[5]

Another conceivable weaker interpretation is to understand metarepresentation as *representation of something that happens to be a representation,*[6] without representing its representational relation to what it represents—falling short of Pylyshyn's definition.[7] A simple and intuitive example would be when thinking of someone thinking, where thinking is simply understood as an activity—that people in the pose of Rodin's *Thinker* are engaged in—without any understanding that this process can only be a process of thinking if the thinker is thinking about something, i.e. that his thinking has representational content. Here too, metacognition in this sense is not recursive because the thought-about thinking has no content that allows for recursive application.

In sum, we can take a narrow-scope interpretation of metacognition in the spirit of the classical trilogy of mental states (Hilgard 1980) as concerning only epistemic (cognitive) mental states or take a wider interpretation including all mental states in the spirit of Cognitive Science. I will stick to the wider usage. We can also take different views on the effects of the prefix 'meta-' and how it produces at least three different wider scope meanings of 'metacognition'.

1. Special Cognition: cognition *beyond* standard cognition (on the way to metacognition):
   a. Vehicle Reference: referring to something that happens to be a cognition.

[3] Minimally only the existence of the relata has to be acknowledged, not any precise form of them, as Leslie and Roth (1993) once thought was being claimed. That is, in the case of mental representation, one need not—and we typically do not—represent the fact that their vehicles are neural states, and do not represent whether the content is in form of linguistic or analogue pictorial form, etc.

[4] Pylyshyn's definition implies an understanding that the represented representation is a representation, which in turn implies some minimal understanding of what a representation is, namely as characterized by its representational relationship to what it represents. This led me to define metarepresentation as *representation of a representation as a representation* (Perner 1991). In analogy, metacognition would consist of cognition about cognitions as cognitions, which implies some understanding of what makes cognitions what they are, namely their representational content.

[5] The case of pretence seems to contradict this conclusion. For, I can pretend that my pretend character is pretending something (metapretence). My claim would be that in this case I need to understand the pretended pretence as a representation, i.e. as something that has content and not just something that is content.

[6] This is similar to what Proust (2010 p. 7; chapter 14 this volume) means with 'de re' in connection with how epistemic emotions refer to the epistemic state.

[7] Perner (1991) gave an illustration in terms of metalinguistics, of illiterate workmen referring to objects in the shape of B, A, and R, which they have to mount atop the entry to a bar, as letters. Although their conversations about these objects refer to linguistic entities (letters) their discourse is 'metalinguistic' only in this weak sense discussed here.

b. Content Reference: referring to the content of a cognition.

2. Recursive Cognition: cognition *about* cognition (as cognition):

c. Representational Reference: cognition about cognition as cognition.

What makes the claim that animals or young children are able of metacognition is the idea that they are able of recursive cognition (cognition about cognition as cognition). For, if it were just an ability to recognize being in a state of thinking (cognition about cognition), without any concern for the intentional content of the thinking, then this ability would be similar to recognizing that one is in a state of digesting—which would also be an interesting 'reflective' ability but not one for which the term 'metacognition' would be most natural.

Nevertheless it is useful to be aware of the looser meanings of 'metacognition' outlined above. They may explain why some behaviour feels intuitively metacognitive even though cognitive analysis shows no need to be based on recursive cognition. As shown in Fig. 6.2 these cases could be considered to lie on the slope from ordinary object-level cognition to full blown recursive metacognition at the higher level. In this sense they can be considered minimally metacognitive or 'minimeta'. Fig. 6.2 also shows additional steps to be discussed later.

A pair of pernicious problems

Inferring metacognitive abilities from behavioural data has proven to be less than straightforward. Here I point out some of the deeper methodological and conceptual problems, which make it very difficult to infer metacognition—understood as recursive cognition—on the basis of behavioural indicators. I intend to illustrate these problems on existing techniques, which have been almost exclusively developed in the comparative animal literature.

I want to emphasize that my goals are to find techniques that can overcome these problems for potential use with pre- or low-verbal children and to draw awareness to these general problems affecting our theory of mental representation. In any case, I do not want to conclude my survey with denying animals any ability for metacognition. However, in checking whether experimental paradigms surpass my problems or not I have to check particular studies. Here details of the studies often matter, and one can be easily accused of 'cherry picking', i.e. picking only on the easy targets. This would be counterproductive for my enterprise; I should pick the hard targets ('inverse cherry picking'). They are more likely to lead me to promising methods. Reviewers of an early draft have already directed me to some tougher targets. I also hope that my arguments will

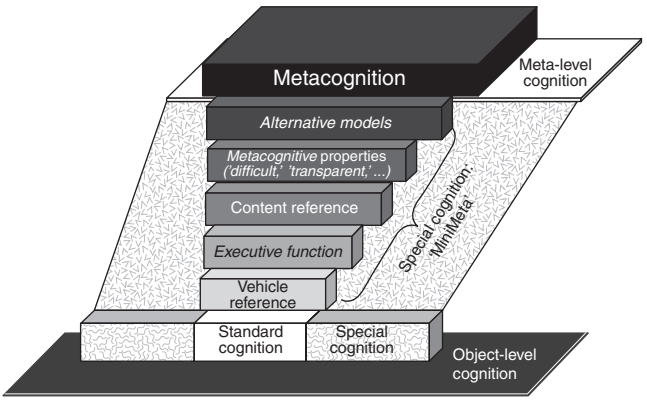


Fig. 6.2 The landscape of object-level and metalevel cognitions.

be taken seriously enough to be controlled for in future research, which is a generally desirable feature (Shettleworth 2010).

BEING in a state versus KNOWING that one is in this state

To see the depth of this problem a brief (I promise!) excursion into why Cognitive Psychology needs to posit mental representations will help. Psychology attempts to explain intelligent or adaptive behaviour, which consists of movements that change the world in the service of reaching a goal by adapting to the given circumstances. So, goals and circumstances determine the movement. For this to be possible there must be some causal link from goals and circumstances to an animal's movement. Since there is no direct physical cause discernible,[8] we need to posit some 'internal' entities that correspond to external situations and goal states, i.e. neural states that represent these external causes. These neural states do actually exist (unlike non-existing goals) and, thus, are able to cause behaviour.

Now we go one step up considering adaptive responses to other sentient beings. If I want to predict whether you will come to our meeting I have to know whether you know that it takes place (a clearly recursive representation). So if I can make a correct prediction we have some basis for claiming that I must have a metarepresentation of your knowledge. Now we go one step in and consider knowledge of one's own knowledge state (the central case of metacognition). If I need to predict whether I will go to the meeting I just have to know that it takes place. I do not need to know that I know it takes place. This would be superfluous. Hence, from my correct prediction of what I will do we cannot convincingly infer metacognition as a recursive concern about my own knowledge.

Whence this asymmetry? Proust (2007) made much of this asymmetry to conclude that metacognition differs essentially from metarepresentation. My explanation is very simple. Your knowledge state cannot take causal influence on my predictions unless I metarepresent it by knowing that you know about the meeting. In contrast, my own knowledge about the meeting is an integral part of my cognitive system involved in generating predictions. It can, therefore, causally effect the prediction of my behaviour directly, without need to metarepresent my knowledge.

One reason why we find it difficult to distinguish *knowing* from the recursive *knowing that one knows* may be our Cartesian intuition, where our mind is considered to be transparent to itself (Churchland 1984). This also explains the reluctance with which the idea of unconscious knowledge (knowledge of which one does not know that one has it) was accepted in scientific psychology. Clearly, with such strong common sense intuitions it is tempting to equate an effect of BEING in a state of knowing with KNOWING that one knows. Yet, keeping these two things distinct is essential. Every behaviour that takes environmental conditions into account depends on knowledge of the environment. If knowing were the same as knowing that one knows then just about every behaviour would be evidence for metacognition.

Now, Call and Carpenter (2001) claimed to have demonstrated metacognitive abilities by showing that chimpanzees as well as very young children reacted differently depending on being knowledgeable or ignorant about the location of a reward. Participants could see one of two tubes being baited. In one condition (full knowledge) they saw which tube this was, in the other (partial ignorance) they did not. Then they had to point to a tube. If it contained the bait they got it, otherwise they did not. Chimpanzees as well as 2½ year old children were able to adapt their behaviour to their knowledge or partial ignorance. They looked first through the tubes to see which one

[8] In particular, goal states do not even exist at the time behaviour is caused. So how could they possibly have causal efficacy.

was baited more often when they were partially ignorant than when they knew where the bait was (superfluous looking). This behavioural difference between the knowledge condition (not looking) and the partial ignorance condition (looking into the tubes) is interpreted as evidence of metacognition (see Call Chapter 4, this volume, for more details on this test and other related tests). How can that be? Chimpanzees, who know where the bait is, go for it and do not look around; when they do not know where it is (total ignorance) they look around; when they know roughly where it is (partial ignorance) they go there and then look around within that region. This behaviour can be governed by the degree of the chimpanzee's knowledge/ignorance without any recursive cognitions about his degree of knowledge. In fact, what would it help the chimp to know that he knows that the bait is in one of the tubes and to know that he does not know in which one? So there is no compelling evidence for recursive cognition (metacognition) in these experiments. One question though is whether the evidence can be firmed up by other features of the chimpanzees' behaviour (see 'Information seeking' section). Another question that remains is why we, or many of us, think or feel that metacognition is involved. This question will be taken up in the penultimate section on MiniMeta.

Internal—external: representing a state of the world versus representing an inner state (representation) caused by the external state

Representing a state of the world is standard object-level cognition. Representing an internal state caused by the external state smacks of potential metacognition. Is this impression warranted?

Let me illustrate the problem with a little thought experiment. When I enter an overheated room I will get hot. This internal physical state is not a cognitive state. It is, however, necessary for giving rise to two kinds of cognitions: feeling hot and realizing that this is a hot room. In either case I might open the window. From my opening the window an external observer could not tell whether I do so because I feel hot (a cognition about my inner state) or because I realized the room is too hot (a cognition about an external state of the world). That is, the observer cannot determine whether my cognitive system is concerned with my inner state or with the state of the room.

This indeterminism, I argue, affects also most experimental demonstrations of metacognition, in particular, those based on the so-called 'opt-out' paradigm. I use Smith et al.'s (1997) classic psychophysical pixel density test (see Couchman et al. Chapter 1, this volume) to illustrate the problem. In their study rhesus monkeys were trained to make one response ('d') when pixel density was dense and make another response ('s') when pixel density was sparse. These response options had *task dependent outcomes*. For correct responses monkeys got a food reward, for incorrect responses they had to endure a time out period. Monkeys learned this task to perfection at the clear density values but made errors close to the border line of sparse and dense. After reasonable mastery a third, *task independent* response option ('o') was introduced. Making this response resulted in a standard outcome (independent of pixel density) of medium value.[9] Both monkeys

[9] Since this payoff structure was used in many studies with quite different kinds of rewards or non-rewards I try to give a more abstract characterization that, in my view, does justice to all variants. When the correct option is taken the subjective expected utility (SEU) of the outcome is high, if the wrong option is taken it is low. Clearly this must be so, because if animals subjectively valued a food pellet less than a time out, they would never learn to respond correctly. Similarly, they must be sensitive to the probability with which responding with 's' or 'd' results in the valued outcome, hence subjective *expected* utility. The critical element of all opt-out paradigms is an additional response option with a *task independent* outcome. I denote this option with 'o' (for opt-out). Its outcome is first of all independent of what the set task is (e.g. the density discrimination) and it must have an SEU below that of the correct response or else the

learned to use the o-response appropriately in the region of objective highest uncertainty, i.e. most often at those density values at which their use of the other two responses was equi-frequent. The authors' 'metacognition' interpretation of this result is that the monkeys learned to use the o-response in response to detecting their own uncertainty or to feeling uncertain. And many have followed suit with this interpretation.

Unfortunately, this interpretation suffers from the same indeterminism as outlined above with 'being hot'. Being uncertain is, like being hot, an inner state. It is, unlike being hot, also a cognitive state—but it is not a metacognitive state (see earlier section 'BEING in a state versus KNOWING that one is in this state'). This inner state is necessary for two types of further cognitions required to guide a learned response: realizing that I am uncertain (a clear metacognition), or realizing that this is a difficult problem (a cognition about an external state of the world). Either one of them can be the basis for my adaptively emitting the o-response. Consequently, an observer of my o-response cannot determine whether my cognitive system is concerned with my inner state of uncertainty or with the difficulty of the task because the behavioural responses afforded to the monkeys are insufficient to distinguish these two possibilities.

A common objection comes with the intuition that judging the difficulty of a task may implicitly already be metacognitive (one could call difficulty a 'metacognitive' property). My Anonymous Reviewer objected to my analysis that *difficult* is a subjective notion and therefore it is hard to see how an animal could assess difficulty without any consideration of its own mental states. It is true, tasks in themselves do not come in degrees of difficulty. Their difficulty depends on how they are managed by the person trying to solve them. But that kind of subjectivity holds for a lot of properties, e.g. relative size (bigger than me, I run; smaller than me, I stay), being hot, or being disgusting (it is disgusting, leave it; it is ok, eat it). Being hot and being disgusting are similar to being difficult in that they depend on evaluation of an external object or event in relation to an internal ability or reaction to it. The maggots on a piece of rotting meat make me feel disgust and I judge the maggots and the whole scene disgusting. Similarly, I read a question on a test, no answer comes to mind, and I judge it difficult and turn the page. If the subjectivity of that judgement is a reliable indicator of metacognition then my quest is at an end: we simply look when children become able to tell when a room feels subjectively hot or when they become able to reject food they deem disgusting. Few would find such a demonstration convincing. So why would learning to skip difficult test items be evidence for metacognition?[10]

Michael Beran in his comments wondered why students' ability to strategically skip multiple choice items is considered a metacognitive skill in all of the traditional metacognition literature in educational science. I surmise that this is so because we have a common sense theory of why test items are difficult. They are difficult because they exceed our knowledge. We tacitly assume that all students share that theory and when they judge an item difficult (not intrinsically metacognitive) they are aware that this is because they don't know the answer (which makes

animals would always choose 'o' (except for erratic exploration). Its SEU must also lie above the average SEU of correct and all incorrect options (randomly choosing any option), or else it would not be chosen on a regular basis. Last but not least, the animal has always the option to not choose any of the above and do nothing (or one of the causally irrelevant activities). Importantly, the SEU of the other responses must be higher than the SEU of doing nothing. I take it that this characterization of the task is a simple logical consequence of the assumption that animals prefer actions with higher SEU over actions with lower SEU outcomes.

[10] The decision to skip can be based on a sense of difficulty not being able to produce anything easily. There is an interesting link to the developmental evidence (Kloo and Rohwer Chapter 10, this volume). Younger children pin their understanding of the word 'know' on being able to give an answer to a question. So they mistake easy guesses for knowing.

it metacognitive). The educational literature is not tuned to asking the foundational issues, which are the focus of this volume, of when one can infer metacognition from task performance. It works under the plausible premise that students metacognitively understand the link between their knowledge and task difficulty, i.e. that they understand or should understand that if the test is difficult then they should have learned harder for it.

Now let me build up a contrasting intuition with an imagined blindsight patient. On each trial we show him in his blind field an X or an O, or sometimes nothing. He is instructed to say 'X' or 'O' or press a button to move to the next trial. When he calls out the same letter as is displayed he receives a nice chime, otherwise a grating scratch, unless he presses the move-on button. From his subjective point of view he makes responses totally unrelated to the display where he can't see anything anyway, but feels a natural tendency to say either 'X' or 'O' (because the unconsciously perceived stimulus primes these responses). However, sometimes this tendency is missing (because nothing is presented and no priming takes place) and then he presses the move-on button. If our patient performs well in this task, can we infer that he is having metacognitions of a minimal kind: 'I *saw* an X' or 'I *knew* there was an X'? Blindsight patients evidently have information (unconscious knowledge) about the stimuli, or else they could not respond contingently. What they lack is awareness of their mental state (metacognition) with which they behold that information. So it seems that from the stimulus-contingent behaviour we cannot, and in this case should not, infer any metacognitive insights.

This undecidability of whether a representation of an inner mental state or of an outer worldly state drives our behaviour is a problem of even wider significance than my examples might suggest: a really pernicious problem. Evans (1982) proposed the notion of an *ascent routine* that enables us to ascend from what we (in our judgement, from our perspective) consider a fact to the mental realm and attribute with logical impunity a corresponding belief to ourselves. For instance, if I take for a *fact* that Obama was born in the USA, then I must *believe* that he was born in the USA. Gordon (1995) extended this idea in the service of simulation theory to all mental states.[11] This correspondence between own mental states and external properties makes it a serious problem of ever knowing the cognitive basis of behaviour that occurs consistently with a particular mental state. Does it depend on recursive cognition about that mental state or on object-level cognition about the external fact from which one can ascend to that mental state?

Summary

I have pointed out two problems for showing metacognition (recursive cognition) on a behavioural basis, and illustrated these problems with examples from classical metacognition experiments. I now go through some of the more recent evidence and show that all (the toughest nuts I could find) founder in their metacognitive claims on at least one of these problems.

[11] Whatever is a fact in my judgement is also (unfailingly) something I believe to be so; whatever is in my judgement frightening (as opposed to in my judgement frightening to others) is also something I am frightened of; whatever is in my judgement desirable is something I desire, etc. A good intuition check for these connections is an extension of Moore's paradox for belief (Moore 1942; Gordon 2007): The claim 'I believe that p and not p' seems contradictory without being a logical contradiction. By extension a similar non-logical contradictoriness seems to adhere to: 'I desire that p and p is (in my judgement) not desirable,' and 'I am frightened by x and x is (in my judgement) not frightening'.

That ascent routines exist for every mental state is a strong claim. A problematic case seems to be 'hope' (Goldman cited by Gordon 2007). I maintain that an ascent routine for hope is possible provided one admits conditional statements about the world. Given that 'it would be a good thing if our CO_2 emissions will be reduced,' I can ascend to attributing to myself: 'I *hope* that CO_2 emissions will be reduced'.

Recursive cognition in animals: which is the best method to detect it?

Information seeking

Call and Carpenter (2001) introduced the information seeking paradigm used for chimpanzees and very young children that I described in the earlier 'BEING in a state versus KNOWING that one is in this state' section. Members of both species differentiated between trials where they knew in which of several tubes a bait was hidden and trials where they only had partial knowledge that the bait was in one of the tubes but not which. They looked through the tubes more frequently in the partial than full knowledge condition in order to find out where the bait was before committing themselves to a final choice.

My analysis concluded that there was no evidence for recursive cognition in these data. When an animal wants food it looks around to find some. When it knows where food is it goes straight there. Nobody would claim metacognition being involved. The tubes experiment only shows that animals can restrict their exploratory behaviour to the region within which they know (partial knowledge) that food can be found.

There are now several follow-up studies. Call (chapter 4 this volume) agrees that random search exploratory behaviour does not need metacognition to be explained. But he argues that metacognition is required to explain the very specific behaviour that animals show in pursuit of knowledge in these new studies. So this is the cherry to pick because it is the toughest nut to crack.

Krachun and Call (2009) showed that animals not only look inside the relevant containers, but they are also very adept at positioning themselves in relation to the differently shaped containers so that they could look inside: 'The crucial aspect of this study was that owing to the containers' diverse geometry and their position on the platform, subjects had to position themselves in different locations depending on the container to spy the food' (Call Chapter 4, this volume). This is an impressive cognitive feat, but how does it relate to metacognition? Complexity of cognition does not make it metacognitive, as Penn and Povinelli (2007) have emphasized in the context of theory of mind. In fact, my worries go beyond the question of whether this sophisticated behaviour needs metacognition to be explained. I wonder whether metacognition could play any helpful role at all. Let me go back to first principles.

An animal's behaviour is guided by a goal and knowledge of how to achieve that goal. So if an animal wants to get to the food it will go where it knows the food is. If there is no such knowledge the desire in combination with its ignorance will trigger search. This can be just random looking around. No metacognition need be involved as Call agrees. If the animal has partial knowledge of where the food is the random search will be constrained by that knowledge. Still no metacognition required.

Now the question is how can we explain animals that do not just look around randomly but who engage in strategically locating themselves in relation to containers in order to get a look inside. I suspect the critical assumption here is that random looking can be understood as a 'response' to not knowing where the food is, while guided looking needs to be understood as instrumental behaviour in order to achieve a goal (getting a good look inside). So we need a goal. Evidently, wanting the food in combination with insufficient knowledge about its location generates a goal of wanting to get a look at the food. The animal must have general knowledge how to achieve this goal. For instance, it has to know that it needs to look inside every corner, niche, or container within the relevant search space and it needs knowledge about how to position itself in order to get a look inside containers, etc.

This seems an explanation without reference to metacognition. Is there a gap in this explanation that needs to be filled by metacognition? One such gap may be the triggering of the new goal to get a look at the food. The claim is that this can be triggered by wanting the food and by not knowing where it is, which are intentional states but not metacognitive states (see section 'BEING in a state versus KNOWING that one is in this state'). In fact, for explaining how lack of knowledge triggers the desire for more information I can see no explanatory advantage of resorting to metacognition. We would have to assume that wanting food and not knowing where it is first leads the animal becoming metacognitively aware that it does not know where the food is, then that metacognition triggers the desire to get a look at the food. Why would that explanation be more elegant or more complete than the original?[12]

There may be another gap residing within my account just given: the goal of getting a good look includes a metacognitive understanding that a good look will provide a particular visual experience. Krachun and Call's (2009) frequent reference to 'visual perspective taking' suggests such a tacit assumption. Certainly, the point of the looking is to get a new visual perspective and with it a new experience and knowledge. But does the animal have to understand this? Is it not enough to aim for a good look rather than a good visual experience? The deeper reason for the good look need not be apparent to the animal, only to evolution, which provided the animal with the desire for getting good looks.

Another critical gap in my argument may be the explanation of how the animal can position itself correctly in relation to a container in order to get a good look inside. For this ability the animal needs first of all intricate knowledge of how to achieve a good look. I cannot see how any metacognitive knowledge can add anything.

In sum, the intricacies of animals' guided search call for a goal to get a good look inside all relevant places in the search space and intricate knowledge about how to achieve such a good look under specific circumstances. How metaknowledge can help here remains mysterious. So my argument is not that an alternative explanation can be found for guided search by cherry picking case specific explanations, but that reference to metacognition does not contribute in any way to an explanation for the animals impressive search behaviour.

Beran and Smith (2011) elaborated on a quite intricate information seeking procedure used unsuccessfully on pigeons before (Roberts et al. 2009). Monkeys were shaped to choose (operating a joystick) one icon of two that revealed a sample stimulus. They then had to choose the other icon, upon which the sample disappeared and three test stimuli appeared. One of them was the same as the earlier sample. For choosing the matching stimulus a reward was given. Both species, rhesus macaques as well as capuchin monkeys, succeeded on this part of the test which pigeons solidly had failed. In a series of further sessions the other three possible variations of occluded/visible sample and occluded/visible test stimuli were successively introduced. The most difficult final trial block was one where all four versions were presented intermixed. Four of eight macaques were able to reach this final phase and learn the optimal response for the last version added, which required to reveal the sample while the test options were already visible (Occluded Sample—Revealed Comparisons; experiment 2). In contrast, not one of seven capuchins managed that level (experiment 3).

Mastery of this condition is intricate. However, the intricacy lies first of all in realizing that one needs to choose the test item that is the same as the sample and for that one needs the sample.

[12] Importantly, I question only how metacognition could possibly improve my explanation, i.e. make it more watertight. I am not questioning that metacognition could improve the animal's way of optimizing its search for more information or other ways of dealing with this situation.

When the sample is or was not present then one needs to produce it by clicking the 'produce sample' icon. But this does not involve any (clearly metacognitive) recursive cognitions like 'Where is *information* about the sample?' or 'Where can I *see* the sample?'. It is not even clear why such cognitions, if animals were capable of them, were of any help to them beyond the basic cognition: 'What is the sample?'.

An interesting question here is why pigeons are apparently incapable of looking for a sample when they are presumably perfectly able to look for food. A plausible reason would be that looking for a sample in order to find a matching item later is of course a greater intellectual challenge than looking for food to consume. Being capable of greater cognitive complexity does, however, not make for metacognition.

Another interesting question concerns the source of the difficulty in the final phase of this study, which some rhesus macaques mastered but no capuchin monkey was able to. Beran (pers. comm.) sees the critical aspect in the need for monitoring: 'Thus, there is a monitoring component to this, not just a goal cognition, because information seeking behaviour (or, in the condition where all information is presented already—the lack of information seeking behaviour) is driven by not just goal cognitions but also assessment of the current environment against that goal cognition'.

I agree with this assessment, but it strikes me that it is a description of the most basic practical reasoning we ascribe to any organism whose movements we call 'behaviour'.

In any case I can see no real need for metacognitive monitoring (monitoring one's mental states over and above monitoring the environment). For instance, in the two conditions in which the test stimuli are visible the animal wants to press the stimulus that matches the sample. If the sample is visible the animal knows what it is and clicks on the matching item. If the sample is not visible the animal generates the subgoal of producing the sample by clicking on the 'get sample' icon. The sample appears, the animal knows what the sample is, and it clicks on the matching test item. Evidently the animal has to go through several cognitive states of wanting to get something, being ignorant of some things and knowing other things. But what help could be provided by additional metacognitions: the animal knowing that it wants to produce the sample, knowing that it does not know what the sample is, knowing that it sees the sample, etc.? In conclusion: there is no convincing evidence from this approach that recursive cognition must be involved.

Call (2010) showed with the tubes set up that great apes (chimpanzees, gorillas, orangutans, and bonobos) checked the tubes more often when partially ignorant than when knowledgeable. In addition (experiment 3: 'passport effect'),[13] the apes checked more often when the bait was particularly attractive (grape) than when it was not (carrot). A small difference of less than 10%, but it occurred under partial ignorance as much as under full knowledge (when apes had seen which tube the reward was put inside). This effect (not the effect of ignorance vs. knowledge) could be due to a simple preference for looking at attractive rewards more than at less attractive ones. So at this point the results are not very telling. But let us assume this preference factor can be excluded, e.g. they know the bait is not for them but for the caretaker. They still check what the caretaker gets but check as often for high- as for low-quality food. In that case the difference in looking would get closer to the theoretical significance that Call was aiming for. The 'superfluous' checking would seem triggered by being afraid of getting it wrong.

However, is the reason for checking really the fear of getting it wrong, or rather fear of it not being there anymore? My intuition about my checking my luggage for my passport is that I check

[13] Josep Call refers to the 'superfluous' checking whether the object is where you know it is as his 'passport effect', because when he travels he packs the passport the night before and then keeps checking every so often to reassure himself that it is still there even though he very well knows it is there.

that I have put it in there. I am not checking whether my assumption (belief) that it is in there may be wrong. Or at least, my intuition is not clear enough to tell between these two reasons. And my intuition as to why the animal checks the food in the tubes is even less clear. Does it check the fact that the bait is indeed in the respective tube or does it check its knowledge of this fact. As the discussion of Internal-external (ascent routines in section 'Internal–external') has made clear this question cannot be easily answered and not on existing data.

Opt-out

I have described the basic opt-out paradigm (version by Smith et al. 1997) in section 'Internal–external' as an illustration of how the internal–external problem affects interpretation of these data. In particular it makes it difficult to tell whether the animal responded to being uncertain or to being faced with a difficult task. I should point out that the basic paradigm is also affected by my other pernicious problem ('BEING in a state versus KNOWING that one is in this state' section). Even if animals do respond to *being uncertain*, it would not be evidence for recursive metacognition, for it would be a response caused by the animal BEING uncertain and not by the animal KNOWING that it is uncertain.

Interestingly, the being-versus-knowing-that-one-is problem takes a slight, but relevantly different role in the opt-out paradigm than in the information seeking case discussed in the 'BEING in a state versus KNOWING that one is in this state' section. Lack of knowledge, presumably—or so I have argued, triggers automatically a desire for engaging in information gathering activity. For this to happen, the animal does not need to know that it lacks sufficient knowledge. This is different in the opt-out experiment where uncertainty (lack of knowledge) does not trigger anything that would lead to an o-response. The o-response has to be learned and become associated with the state of being uncertain. Perhaps one could argue that mental states like uncertainty can only cause innately specified effects (looking around randomly) but an animal cannot learn to associate a novel behaviour with it directly. This can only happen when the animal is aware of being uncertain.

I do not know whether there is anything to this idea; but it is interesting. So let me assume for argument's sake that conditioning of a response to an inner state requires awareness (knowledge) of being in that state. Davidson (1987, 1993) has looked at conditioning of fear responses to inner states of food deprivation in rats. For instance, rats were food deprived for either 23 or only 6 hours, and then received a shock depending on condition after 23 or after 6 hours deprivation. This shock was always at least 6 hours after last food intake to prevent conditioning to recent memory of food. After 24 trials rats differentiated between deprivation levels by their differential rate of freezing in anticipation of the shock. Then to test whether freezing was conditioned to the inner state of food deprivation rather than external stimulus aspects rats were either given an intubation of high-calorie food or a sham intubation. Relative to sham intubation, the high-calorie load increased freezing for rats previously shocked under 6-hour food deprivation and decreased freezing for rats previously shocked under 23-hour food deprivation. So, clearly freezing in anticipation of shock can be conditioned to the inner state of being hungry. If, by hypothesis, conditioning to an inner state requires knowledge of the inner state then these results show that rats can be aware of their state of hunger. If, in contrast, we agree that behaviour can be conditioned directly to an inner state (without the animal knowing that it is in that state) then no reflective abilities can be claimed for rats. But then the findings from the opt-out experiments that training animals to use the o-response when being uncertain do not seem to require any metacognitive ability either.

The basic opt-out paradigm as described in its version used by Smith et al (1997) has undergone great evolution since its conception. So the hope arises that the more refined recent versions might overcome my pernicious problems pair.

Smith et al. (2006) introduced deferred feedback to 'make it impossible for the uncertainty response [i.e., o-response] to be conditioned by feedback signals, responsive to reinforcement history, or based in low-level associative cues' (p. 289). The o-response was trained in groups of four trials consisting of a mix of easy and difficult problems. The animal's response and deserved outcome was recorded but the outcome for each trial was only paid out after each block of four trials: first all the rewards for correct responses were paid out and then the sum of penalty timeouts for wrong responses had to be endured. The o-response counted neither for rewards nor for a time out. Smith et al. (2008) showed that modelling animals' performance with an associative model assuming a baseline tendency to use the o-response provided a good fit to the data under transparent feedback but not to data gained under deferred feedback. It was concluded that animals must have used their state of uncertainty for learning when to use the o-response. However this conclusion still leaves open whether the o-response was conditioned to KNOWING that they were uncertain (metacognition) or just to BEING uncertain (no metacognition: Problem of section 'BEING in a state versus KNOWING that one is in this state') and whether it was conditioned as a response to being uncertain (a mental state) or to encountering a difficult trial (an external condition).

Washburn et al. (2006) showed that use of the o-response transfers immediately to novel discriminations (see also Kornell et al. 2007; Couchman et al. 2010). Such transfer provides important evidence that the o-response could not have been conditioned to the middle range of the discriminative (density) stimulus dimension (e.g. if semi-dense, take o-response; known as the 'middle stimulus argument'). Unfortunately transfer to novel dimensions does not help solve my pernicious pair of interpretation problems. The transfer can be explained by both, conditioning to the state of being uncertain ('BEING in a state vs. KNOWING that one is in this state') or to difficult trials ('Internal–external: representing a state of the world versus representing an inner state (representation) caused by the external state'). It is hard to see what a metacognitive account of animals being conditioned to knowing that they are uncertain would add.

Hampton (2001) introduced the memory version of the basic opt-out paradigm. Monkeys were first presented with a sample, and then after a delay with a choice between opting-out (low reward: pellet) or opting-in followed by an array of stimuli containing one identical to the initial sample. Correct choice of the stimulus that matched the sample led to high reward (peanut), error choice to a 15-second timeout.

Again this can be explained by conditioning to having a memory without need for monkeys to be aware of the quality of their memory as the metacognitive interpretation has it. What monkeys learn is that if there was no sample then they should opt out, otherwise they should opt in and subsequently press the icon corresponding to the sample. Since there was always a sample but the monkey's knowledge of it depended on its memory, the longer the delay between sample and test the more likely that no memory could be retrieved. The important point is that the monkey need not monitor the quality of its memorial process (a metacognitive act) but only monitor what had happened in the past, which is presented through memorial processes.

Washburn et al. (2010) used transcranial magnetic stimulation (TMS) to interfere with a monkey's early processing of visual images. The monkey's recognition accuracy declined and its use of the o-response increased. Again this result does not help decide whether the monkey was monitoring the past and used the o-response when—from its point of view—there had been no sample stimulus or whether it was monitoring the quality of its memory and used the o-response when judging that its memory has faded or gone.

Kornell et al. (2007) introduced yet another twist to this paradigm: the retrospective opt-out (confidence judgements) and combined it with the transfer to completely novel tasks. Two rhesus macaques were trained to indicate the large dot in a field of many slightly smaller dots. Difficulty of

the task depended on the difference in size. After one of the dots had been indicated two icons appeared: HR (high risk) and LR (low risk). Choosing the HR-icon resulted in a large reward (+3 tokens) if the large (correct) dot had been indicated beforehand, but resulted in a loss (−3 tokens) if any other (wrong) dot had been indicated. If the LR-icon was chosen a small reward (+1 token) accrued regardless of what had gone before. Monkeys learned to press the HR icon after easy trials and the LR icon after difficult trials where they were more likely to have chosen a wrong dot. The metacognitive interpretation of these data is that monkeys realized when they might have made a mistake and learned to signal their uncertainty or low confidence by choosing the LR icon. However, I can think of at least three alternative explanations that do not bring metacognition into play.

(1) Monkeys' behaviour in this basic task can be explained by assuming they learn to press the largest dot if they detect one, otherwise they pick any dot. The likelihood that they detect a largest dot is higher for easy trials, where the largest dot is easy to see, than for difficult trials. They also learn that after having chosen the largest dot, provided there was one different enough for them to see, it pays to follow this choice by choosing the HR icon; but if—from their point of view—there was not a largest dot to press (or there was but they chose the wrong dot) they should follow it up by choosing the LR icon. Problems for this explanation arise from the transfer findings. After the dot discrimination task the monkeys were trained on a recognition memory task. After a sequence of about 6 pictures a test slide showing nine pictures simultaneously was presented. Animals had to choose the picture that had been presented before in the sequence of pictures. After mastery of this memory task the familiar LR- and HR-icons were presented after each choice of test picture. Amazingly, right from the beginning (first 4 test sessions) the animals showed a significant tendency to follow correct selection of the old picture with choice of the HR-icon and wrong choices of new test items with the LR-icon. This result would not be possible under my explanation. Since there was no largest dot to press in the memory task they should always opt for the LR icon.

This problem for my explanation can be rectified by assuming that stimuli promising high payoff are imbued with a value (the largest dot is a good thing). And they get conditioned to choosing the HR icon after having chosen a good thing, the LR icon otherwise. Under this description of the stimuli the response options can easily transfer to the memory task. In the memory task they learn that the test picture that was the same as the one that occurred in the series before is a really good thing. When you choose it you get a large reward. So this makes them prefer to press HR after choosing the good item and LR if they could not find a good item among the array. The assumption that animals tend to rely more on the appetitive value of stimuli (good, bad) rather than their physical description (largest dot, same as sample, etc.) is underlined by the standard position on ultra quick processing of the emotional significance of stimuli (LeDoux 1996; Zajonc 1980).

There are two more alternative explanations that keep closer to the metacognitive interpretation. (2) The choice of LR depends on being uncertain but it is the state of uncertainty that makes the animal prefer the LR icon and not a metacognitive process of the animal realizing that it is uncertain. This approach will directly transfer to the memory task. (3) The animal distinguishes between easy and difficult trials, chooses to follow easy trials with HR and difficult trials with LR, and does the same in the first discrimination task and later in the memory task. None of these explanations requires clearly recursive metacognition.

Kiani and Shadlen (2009) taught rhesus monkeys to move their gaze to a target to the right or to the left depending on a left/right movement stimulus at fixation point. One of the target points was presented in the receptive field of neurons in lateral interparietal (LIP) cortex whose activity was recorded. Task difficulty was varied by the clarity of the movement stimulus. On some trials

an 'opt-out target was shown above fixation point about ½ sec after extinction of the motion. Then the fixation cross extinguished telling the monkey to move his eyes to one of the targets. Looking at a target in the direction of motion led to full drink reward, looking at the opposite target led to a time out, and looking at the opt-out target to 80% of drink reward'.[14] As expected, accuracy went up and frequency of looking at the opt-out target went down with clarity of the motion stimulus. What is new is that the LIP neurons not only predicted at which of the presented targets (in direction of motion or opposite) the animal would later look. They also predicted whether the animal would look at the opt-out target before it even knew whether such a target would be available or not. The authors present a model to explain their data and mention that their model makes metacognitive explanations for certainty monitoring unnecessary (Kiani and Shadlen 2009, p. 763). What I would like to know is what the neural response had to be so that a metacognitive explanation would be needed.

First- to third-person perspective transfer: awareness of seeing

There is an interesting methodological approach in theory of mind to showing that animals and infants understand other people's looking as a mental state of seeing. Its core argument draws on the involvement of metacognition. For instance, experiments by Hare et al. (2000) show that chimpanzees can anticipate a conspecific's likely action depending on whether the other can look at the bait behind a transparent screen or cannot look at it because it is shielded by an opaque screen. The question of theoretical significance is what this shows about their understanding of the mind. The 'mentalist' explanation (e.g. Tomasello, Call, and Hare 2003) assumes that chimpanzees really understand that *looking* at the object in its hiding place (looking being a purely observable external event) leads to *seeing* the object and where it is (a mental state with subjective content), which leads to *knowing*[15] where the object is (a mental state that enables to adapt behaviour to the observed event). The alternative proposal (Povinelli and Vonk 2003; Penn and Povinelli 2007; Perner 2010) refers to 'behaviour rules' that inferentially link observable behaviour (looking) with observable behaviour (go for the food) without a mediating chain of inferences involving mental states like seeing and knowing. That is, chimpanzees understand that if a conspecific looks or has looked at the place with food then he will go for it when able to do so.

One longstanding proposal for deciding this issue is to use a method pioneered by Novey (1975) with infants to test for their ability to infer from their first person experience with transparent and opaque goggles that another person can or cannot see when wearing these goggles (Heyes 1998). The central idea behind the proposed investigation is that there could be no behaviour rule relating the personal experience of seeing things when wearing these goggles to the experience of seeing another person wearing these goggles and directing his head towards a target object or event. Meltzoff and Brooks (2008) found that 1-year-old infants can use this information and are more prone to follow an adult's head direction with their gaze when the adult wears a blindfold that they have experienced as being transparent than when wearing one they had experienced as opaque. Teufel et al. (2010) used goggles and reported that by 2½ years children can also verbally indicate through which goggles the other could see and through which he could not. This suggests—so the argument goes—that such young infants must be aware, not just of what they see, but of the fact that they can see, a clearly metacognitive awareness, or else it would be inconceivable how they could possibly make a link between their experience of seeing/not

[14] Smith et al (2008) would not count this study as a serious contender because by rewarding the use of an opt-out response opens the gate for behaviourist alternative explanations.

[15] Tomasello and Call (2006) changed their position on this second claim.

seeing with the goggles to what another person can or cannot experience when wearing these goggles.

This procedure has been unsuccessfully used with chimpanzees.[16] Before someone rushes into improving the technique for a repeat, a word of caution: these data with infants may not show what people take them to show. We have known for some time that chimpanzees can distinguish between transparent and opaque screens (see results from Hare et al. 2000) and they seem to be particularly sophisticated at understanding what transparency/opaqueness affords (see the informal observations photographically documented in Povinelli 1996). No one has claimed that this is evidence for metacognition in chimpanzees. Going along with this intuition I conclude that the ability to distinguish transparent from opaque screens in infants is no evidence for metacognition either. We need to ask how we determine whether something is transparent or opaque. Presumably in the same way we determine whether something is red or blue, or square or round, by using our own personal experience with it: if it looks blue it is blue . . . if it looks transparent it is transparent.

The goggles and blindfolds used in the infant experiments differ from the transparent/opaque screens used with chimps in that one cannot see from a distance what they are—one has to bring them close to one's eyes. As a consequence, when observing another person wearing the goggles one cannot simultaneously see whether they are transparent or opaque. One can only know from memory. But apart from this difference the studies with goggles require the same inferences as those with screens. That is they do not exclude behaviour rules. One has to look through the goggles to determine whether they are translucent or opaque as much as one has to look through the screen to determine whether it is transparent or opaque. Once one knows this one can anticipate the other person's actions and abilities on the basis of behaviour rules: if there is a transparent object (goggles or screen) between his eyes and the target he will behave adaptively towards the target, otherwise not.

Interestingly, this further step does not come so easily to 2½-year-old children as shown in the third experiment by Teufel et al. (2010). Although they could indicate correctly which goggles one could see through, they showed no sign of understanding that transparency made a difference to the wearer's knowledge. While the adult was wearing the glasses the child saw a sticker being put inside one of two containers. The children requesting help opening the container made as many pointing gestures to the container regardless of which goggles the adult had been wearing during hiding. Only when children were given direct experience of the adult being unable to act sensibly when wearing the opaque glasses did they adjust their requesting behaviour accordingly (experiment 3).

So there is no mileage in goggles for deciding the issue plaguing theory of mind in infants and, thus, no mileage for demonstrating metacognition of perception, unless one wants to claim that the ability to distinguish transparent from opaque screens itself requires metacognition. In that case the evidence is already in from chimpanzees' sophistication with transparent objects and no goggle experiment is needed.

MiniMeta

The project

My analysis of some of the most impressive 'metacognition' studies with animals leaves me without a single clear behavioural test, from which one can infer metacognition in young children without having to rely on sophisticated language use. This makes me wonder why many people,

[16] Vonk et al. 2005, unpublished work; manuscript available on request; cited by Penn and Povinelli 2007.

me included, have the initially unquestioned impression that each of these paradigms do demonstrate metacognition[17]—until one engages in a strict cognitive analysis of the phenomena. Of course, one answer to this question may be that our intuition is simply misguided by uncritical application of our folk psychology. The mistake is to think that the behaviour shown in these tasks can only occur for the reasons that we would give when asked to justify or explain our own behaviour in those situations. This tendency is reinforced by our awareness of the studies' objective to assess metacognition, and so our intuitive understanding of the tasks is already framed in metacognitive terms.

Yet, there may be a more objective fact underlying our intuitions. The distinction between object-level and metalevel may not be as dichotomous as these labels suggest. There may be a more continuous slope leading from the lower to the higher level (see Fig. 6.2). Our tendency to see metacognition in 'metacognition' tasks stems from the fact that they require cognitions that are not needed for ordinary object-level cognition but are typical ingredients of metacognition. One could say that they are cognitions that go *beyond* (one meaning of meta-) ordinary cognitions in the direction of full blown recursive metacognition.

The point of this enterprise is not primarily to explain our intuitions about applying the term 'meta-' but to learn more about the nature of metacognitive tasks and what animals and young preverbal children can do in this direction. To safeguard against panmetacognition (seeing metacognition everywhere) I want to adhere to the following three *MiniMeta Check Criteria*:

1. Necessity: is the component cognition that makes behaviour intuitively 'meta-' necessary for the behaviour to occur?

   Demonstrations of metacognition in non-verbal creatures are based on showing behaviour under conditions in which the behaviour could allegedly not be shown unless some special cognitive (minimally metacognitive) processes were involved. So we need to look first whether the behaviour could occur with only patently ordinary cognitive processes. Only if the special cognition turns out to be necessary for this behaviour can further checks provide evidence for MiniMetaCognition.

2. Directionality: MiniMetacognition is cognition that goes beyond ordinary object-level cognition in the direction of recursive metacognition.

   This criterion should ensure that not just any unusual cognition be classified as MiniMeta. Only cognitions that have some affinity with standard metarepresentational metacognition should qualify. For instance, Carruthers and Ritchie (Chapter 5, this volume) questioned whether opt out tasks testing for knowledge of uncertainty require metarepresentational understanding of uncertainty and suggested that they may require a certain feeling generated by uncertainty. Although the queasiness caused by indecision may not be proper recursive cognition of feeling queasy about one's uncertainty, even the first-order state of simply feeling queasy has affinity with queasiness about uncertainty because it was caused by uncertainty.

3. Exclusivity[18]: MiniMetacognition should only be needed for behaviour that is intuitively metacognitive. It should not as well be needed for behaviour that has never been claimed to

[17] I follow here a certain Principle of Understanding: 'Never think you've understood something unless you've also got a good explanation for why others before you kept getting it wrong!' (Perner 1991, footnote p. 58).

[18] This could also be dubbed the 'why more experiments?' argument. If the analysis turns up that long known findings provide equally good evidence of metacognition, then the question arises: Why those additional metacognition experiments?

demonstrate metacognition (unless the new discovery leads to a convincing re-evaluation of the implications of the original findings).

For instance, Proust (2007) has argued that regulatory processes of monitoring and control should count as metacognitive. One could counter that established models of even the most simple action control posit that a corollary discharge from the motor command is used to project the intended movement (forward model), which is then compared with somatosensori feedback of the actual movement, and any registered deviations are used to correct the future movement path in advance (Wolpert et al., 1995). Since such monitoring and control is so common it would make just about all behaviour metacognitive (Carruthers and Ritchie Chapter 5, this volume).

In the following subsections I illustrate the MiniMeta approach with but one fully argued example but indicate other intuitively promising venues.

Implicit awareness of ignorance: information search under partial ignorance

In the section 'BEING in a state versus KNOWING that one is in this state', I described the experiment by Call and Carpenter (2001). Chimpanzees looked more often inside a tube to check where the bait was before committing themselves to a definite choice if they only knew that a bait was in one of the tubes but not which (partial knowledge) than when they knew the precise tube (full knowledge). This has been taken as evidence for metacognition. To see whether this interpretation is warranted I contrasted this result to the typical exploratory behaviour shown by an animal that does not know where the food is (no knowledge). One can accommodate this finding easily within regular object-level cognition: content of knowledge determines behaviour. If the animal knows that the bait is in location x it will retrieve it from x. If it knows it is in location y it will retrieve it from y. If the animal knows nothing about the bait's location it will engage in exploratory behaviour, looking around. So far, there is no intuition that metacognition would be involved. Now, if the animal knows that the bait is in one of the tubes, but not in which one, object-level cognition would lead the animal to restrict its exploration to the tubes. So, why does the partial ignorance case raise the spectre of metacognition?

Here is one MiniMeta idea. In the full knowledge and no knowledge case there either is information about the bait's location or there is no information about it. Whereas in the partial ignorance case the animal has to represent a disjunctive state of affairs: the bait is either in tube 1 or in tube 2. Ordinary cognition in animals uses perceptual input to keep an updated mental model of where things are. A disjunction requires alternative mental models and is an 'implicit' way of representing one's ignorance about the actual location. This would provide an objective feature for our intuition that pointed search under partial knowledge involves metacognition. Now let us see whether this suggestion passes our checks for MiniMeta.

1. Necessity: are alternative models of reality necessary for showing the observed behavioural differences in the knowledge and partial ignorance task? We do not know. The animal might just have a single model that specifies the region of the tubes and, consequently, the animal explores the tubes and not anywhere else. Call and Carpenter (2001) noticed that some individuals used the particularly efficient search of seeing an empty tube and then choosing the other tube without looking inside it first (see also Call, this volume). This speaks for alternative models because if one model is ruled out then only one alternative is left and needs no checking. Whereas, if search is limited to a region then the animal would tend to search until it caught a glimpse of the bait.

The recent evidence by Krachun and Call (2009, see section 'BEING in a state versus KNOWING that one is in this state') that animals place themselves adaptively into the right position for looking inside relevant containers does, however, not help decide the issue. The adaptive self placement requires sophisticated knowledge of how to best explore a container, but it does not require a disjunctive representation 'it could be in one of these containers'. The animal just looks inside every container within the restricted search space.

Nevertheless, although we have no clear evidence for the animals entertaining alternative models, we have a promising line of research to pursue.

2. Directionality: entertaining alternative models is not just any kind of unusual cognition. It clearly points to uncertainty. Hence, although it is not metacognitive in the metarepresentational sense (i.e. the animal knowing that it does not know in which tube the bait is), knowing that the bait could be either here or there is a clear step in that direction (an 'implicit' admission of ignorance).
3. Exclusivity: our MiniMeta evidence consists of animals engaging adaptively in information seeking behaviour because they represent alternative models (implicit knowledge of their partial ignorance). To my limited knowledge of the animal literature I think that there is no evidence that animals can do this in other situations where one would not get the intuitive impression of metacognition being involved.

In conclusion, experiments to show that animals base their information seeking behaviour on alternative models of reality are worth their money. Beck et al. (chapter 11, this volume) use a similar approach to children's understanding of uncertainty.

'Metacognitive' properties

In the section 'Internal–external' I pointed out that many types of mental states are caused by particular types of situations or objects: disgust is caused by disgusting objects, uncertainty is caused by difficult problems, etc. This makes it difficult to decide whether behaviour shown under these conditions is indicative of metacognition about the type of mental state or of plain cognition about the type of situation. There is, though, a strong intuition that these cases differ. For instance, when judging that something is disgusting we would not feel that this requires a metacognitive concern about one's feeling disgusted by the object. In contrast, when a task is judged difficult because one does not know the answer, one does have the feeling that this judgement involves some metacognitive awareness of one's lack of knowledge.[19]

Another property of objects that smacks of 'metacognition' is transparency (see section 'First- to third-person perspective transfer: awareness of seeing'). Yes, I can classify objects as transparent or opaque, but to judge something as transparent don't I need some metacognitive awareness of the fact that I can *see* through it? That is the intuition, but we do not know what really goes into such a judgement. Finding the critical difference between judging something as *disgusting* and judging something as *difficult* or *transparent* would be an important advance for the MiniMeta project. All I can contribute here is to point out the difficulties.

Notions like 'subjectivity' and 'involvement of mental states' (as my Anonymous Reviewer suggested) do not go far enough. Take for example size constancy in visual perception. The task is to judge whether two objects presented at different distances from the observer are the same or different in size. Our visual system can be fooled but by and large can do this quite accurately. When we try to explicate how it can do this we get quickly entrenched in very metacognitively sounding

[19] An intuition I share with my Anonymous Reviewer.

arguments: The visual system *knows* that a more distant object *projects* a smaller retinal image than the same size closer object. The system *takes this into account* by *judging* the size of objects in relation to their *estimated* distance. Does this make *size* a metacognitive property because its judgement relies on knowledge about perception? The strong intuition is that it is not. But then, what makes properties like *difficulty* and *transparency* minimally metacognitive?

Conclusion

I have set out in search of methods for investigating metacognition in children that do not rely on language. I formulated two rather pernicious problems for any attempt to infer metacognition from behavioural data. Unfortunately, search of the rich repertory of methods developed in the comparative animal literature showed that no single method—of those that I thought would be most promising—could overcome these two problems.

Other interpretation problems with these tasks also were discussed. So, the comparative literature has increasingly taken to arguing with cross species consistency (Couchman et al. 2012). Species which show information seeking also learn the opt-out response, like chimpanzees and rhesus macaques, while others fail to show these behaviours, such as pigeons, rats, and capuchin monkeys. The split between these groups of species is not completely clear cut. For instance, capuchins can learn to ask for information which pigeons cannot (Beran and Smith 2011; also see 'Information seeking' section). Nevertheless, this separation into two groups is seen as confirmation of the view that one group is capable of metacognition and the other is not, because the intended common denominator of the different tasks is supposed to be metacognition. Extending this approach to developmental investigations with children is costly. It demands employment of several different tasks and hope for a cross age consistency. And then there is still the possibility that the common feature of the so-called metacognition tasks that separates groups of species is not metacognition but some other kind of cognitive complexity.

Another promising indirect way of strengthening the evidence for metacognition was recently reported by David Smith and his colleagues (Smith 2011). They used their pixel density opt-out task together with a concurrent executive task. Their reasoning was that metacognition makes executive demands and, therefore, a concurrent executive task should interfere with the opt-out response but not with the primary choices (sparse/dense). Indeed with human participants a concurrent variation of the number Stroop task interfered specifically with the opt-out response to the degree that it eliminated the opt-out response altogether. They now also have data from monkeys. A concurrent delayed matching-to-sample task interferes to some degree with the opt-out response but not with the primary response options. These are impressive confirmations of a risky prediction from the metacognition stance.

Yet, it still leaves us with the not implausible possibility that the hallmark of the opt-out response is an executive demand and not necessarily metacognition. As a matter of wild speculation this executive command could be the ability to distance oneself from rash responding. That makes animals which have this ability more likely to gather more information before responding, and not take blind risks with the potentially best paying response. Rhesus macaques and great apes have this ability, and perhaps capuchins, pigeons, and rats lack it.

Although the methods used with animals do not provide unambiguous evidence for metacognition under cognitivist scrutiny, intuitively these tasks do feel metacognitive. One resolution of this contradiction would be if these tasks require abilities that lie between object-level and meta-level cognitions, i.e. they are 'minimally metacognitive'. This would explain the intuition and also why the cognitivist analysis does not admit them as metacognitive in the sense of recursive cognition. In my MiniMeta programme I gave some guidelines of how to identify whether a task

is minimeta. I only managed to outline the strategy and, thus, only presented information search under partial knowledge as a fully argued example. I came to consider the ability to entertain alternative mental models as an implicit way of acknowledging partial ignorance. I also looked at 'metacognitive' properties, properties that can only be detected with the use of metacognition. I was unable to pursue the potential of this line lacking the relevant insights how to proceed. There are very likely many other minimeta lines to explore. A particularly interesting case of cognition between object- and metalevel may be the case of conditional reasoning as elaborated by Johannes Leitgeb in Chapter 15.

So, in the end I cannot be of much help to my colleagues contributing to the next section. My search for the most telling paradigm across the rich repertory of methods in the comparative animal literature did not come up with a clearly satisfactory result. The best suggestion I can pass on is to look for developmental consistency across different 'metacognition' paradigms. Unfortunately this is a costly approach. Beyond that I hope that my pair of pernicious problems will provide a good measurement bar against which to assess the developmental findings.

Acknowledgements

I thank my friends and colleagues who have given me critical pieces of advice and criticism on an earlier version of this chapter: Michael Beran, Johannes Brandl, Josep Call, Tony Dickinson, Frank Esken, Ulrike Klossek, Sara Shettleworth, and last but not least The Anonymous Reviewer. Financial support came from the European Science Foundation and the Austrian Science Fund (FWF Project I93-G15) 'Metacognition of Perspective Differences'.

References

Beran, M. J. and Smith, J. D. (2011). Information seeking by rhesus monkeys (*Macaca mulatta*) and capuchin monkeys (*Cebus apella*). *Cognition*, 120, 90–105.

Call, J. (2010). Do apes know that they can be wrong? *Animal Cognition*, 13, 689–700.

Call, J. and Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition*, 4, 207–20.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language*, 23, 58–89.

Churchland, P. M. (1984). *Matter and consciousness: A contemporary introduction to the philosophy of mind*. Cambridge, MA: MIT Press. A Bradford book.

Couchman, J., Beran, M., Coutinho, M., *et al.* (2012). Do actions speak louder than words? A comparative perspective on implicit vs. explicit metacognition and theory of mind. *British Journal of Developmental Psychology*, 30(Pt 1), 210–21.

Davidson, T. L. (1987). Learning about deprivation intensity stimuli. *Behavioral Neuroscience*, 101, 198–208.

Davidson, T. L. (1993). The nature and function of interoceptive signals to feed: Toward integration of physiological and learning perspectives. *Psychological Review*, 100, 640–57.

Dunlosky, J. and Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage Publications.

Evans, G. (1982). *The varieties of reference*. Oxford: Clarendon Press.

Flavell, J. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34, 906–11.

Fodor, J. A. (1978). Propositional attitudes. *The Monist*, 61, 501–23.

Goodman, N. (1976). *Languages of art*. Indianapolis, IN: Hackett Publishing Co.

Gordon, R. M. (1995). Simulation without introspection or inference from me to you. In M. Davies, and T. Stone (Eds.) *Mental Simulation: Evaluations and applications*, pp. 53–67. Oxford: Blackwell.

Gordon, R. M. (2007). Ascent routines for propositional attitudes. *Synthese*, 159, 151–65.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 5359–62.

Hare, B., Call, J., Agnetta, B., and Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, 59, 771–85.

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, 21, 101–48.

Hilgard, E. R. (1980). The trilogy of mind: Cognition, affection, and conation. *Journal of the History of the Behavioral Sciences*, 16, 107–17.

Kiani, R. and Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 324(5928), 759–64.

Kornell, N., Son, L. K., and Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18, 64–71.

Krachun, C. and Call, J. (2009). Chimpanzees (*Pan troglodytes*) know what can be seen from where. *Animal Cognition*, 12, 317–31.

LeDoux J. E. (Ed.) (1996). *The emotional brain*. New York: Simon & Schuster.

Leslie, A. M. and Roth, D. (1993). What autism teaches us about metarepresentation. In S. Baron-Cohen, H. Tager-Flusberg and D. Cohen (Eds.) *Understanding other minds: Perspectives from autism*, pp. 83–111. Oxford: Oxford University Press.

Meltzoff, A. N. and Brooks, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology*, 44, 1257–65.

Metcalfe, J. (2008). Evolution of metacognition. In J. Dunkosky, and R. A. Bjork (Eds.) *Handbook of metamemory and memory*, pp. 29–46. New York: Psychology Press.

Moore, G. E. (1942). Reply to critics. In P. A. Schilpp (Ed.) *The philosophy of G. E. Moore*. Evanston, IL: Northwestern University Press.

Novey, M. S. (1975). The development of knowledge of other's ability to see. Unpublished Doctoral dissertation, Department of Psychology and Social Relations, Harvard University.

Penn, D. C. and Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society of London, Series B*, 362, 731–44.

Perner, J. (2010). Who took the cog out of cognitive science?—Mentalism in an era of anti-cognitivism. In P. A. Frensch, and R. Schwarzer (Eds.) *Cognition and Neuropsychology: International Perspectives on Psychological Science (Volume 1)*, pp. 241–61. London: Psychology Press.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press. A Bradford book.

Povinelli, D. J. (1996). Chimpazee theory of mind?: the long road to strong inference. In P. Carruthers, and P. K. Smith (Eds.), *Theories of theories of mind*, pp. 293–329. Cambridge: Cambridge University Press.

Povinelli, D. J. and Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, 7, 157–60.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 2, 271–95.

Proust, J. (2010). Metacognition. *Philosophy Compass*, 5, 989–98.

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? *The Behavioral and Brain Sciences*, 1, 592–3.

Roberts, W. A., Feeney, M. C., McMillan, N., MacPherson, K., Musolino, E., and Petter, M. (2009). Do pigeons (*Columba livia*) study for a test? *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 129–42.

Shettleworth, S. J. (2010). Clever animals and killjoy explanations in comparative psychology. *Trends in Cognitive Sciences*, 14(11), 477–81.

Smith, J. D. (2011). Animal Metacognition. Invited contribution to ESF organised symposium 'Thinking About Thinking—How Do We Know What We Know?' at the 2011 AAAS Annual Meeting in Washington, DC, February 17–21.

Smith, J. D., Beran, M. J., Redford, J. S., and Washburn, D. A. (2006). Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General*, 135, 282–97.

Smith, J. D., Shields, W. E., Schull, J., and Washburn, D. A. (1997). The uncertain response in humans and animals. *Cognition*, 62, 75–97.

Smith, J. D., Beran, M. J., Couchman, J. J., and Coutinho, M. V. C. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin and Review*, 15, 679–91.

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In D. Sperber (Ed.) *Metarepresentations. A multidisciplinary perspective*, pp. 117–137. Oxford: Oxford University Press.

Teufel, C., Clayton, N. S., and Russell, J. (2010). Two-year-old children's understanding of visual perception and knowledge formation in others. Unpublished manuscript, Brain Mapping Unit, Department of Psychiatry; University of Cambridge, UK.

Tomasello, M. and Call, J. (2006). Do chimpanzees know what others see—or only what they are looking at? In S. Hurley, and M. Nudds (Eds.), *Rational animals*, pp. 371–84. Oxford: Oxford University Press.

Tomasello, M., Call, J., and Hare, B. (2003). Chimpanzees versus humans: It's not that simple. *Trends in Cognitive Sciences*, 7, 239–40.

Washburn, D. A., Smith, J. D., and Shields, W. E. (2006). Rhesus Monkeys (*Macaca mulatta*) immediately generalize the uncertain response. *Journal of Experimental Psychology: Animal Behavior Processes*, 32, 185–9.

Washburn, D. A., Gulledge, J. P., Beran, M. J., and Smith, J. D. (2010). With his memory magnetically erased, a monkey knows he is uncertain. *Biology Letters*, 6, 160–2.

van Inwagen, P. (2010). Metaphysics. In Zala, E.N. (Ed.). *Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, CA 94305, USA.

Wimmer, H. and Perner, J. (1979). *Kognitionspsychologie*. Stuttgart: Kohlhammer.

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269, 1880–2.

Zajonc, R. (1980). Feeling and thinking: Preferences need no inferences, *American Psychologist*. 35, 151–75.

Section II

Metacognition in human development

This page intentionally left blank

Chapter 7

Metacognition in infants and young children

Beate Sodian, Claudia Thoermer, Susanne Kristen, and Hannah Perst

Introduction

Metacognition has been defined as any knowledge or cognitive activity that takes as its cognitive object, or that regulates, any aspect of any cognitive activity (Flavell et al. 1993, p. 150). Metacognitive knowledge includes knowledge about one's own information processing, as well as knowledge about the nature of cognitive tasks, and about strategies for coping with such tasks. Metacognitive regulation requires executive skills related to monitoring and self-regulation of one's own cognitive activities (Schneider 2008). Flavell (1979) distinguished between three major components of metacognition: metacognitive knowledge, metacognitive experiences, and metacognitive skills, that is, strategies controlling cognition.

Declarative metacognitive knowledge has been studied in children as young as 4 or 5 years in interview studies of metamemory (Kreutzer et al. 1975), as well as in tasks requiring a judgement of the difficulty of a memory problem (Wellman 1978) or a strategy choice (Sodian and Schneider 1990). Even preschoolers possess some factual knowledge about person, task, and strategy variables affecting memory performance, and this knowledge increases rapidly in the kindergarten and elementary school years. Similarly, a basic understanding of mental verbs such as 'know', 'guess', and 'remember' develops in preschool age (Johnson and Wellman 1980).

Studies of procedural metacognition have found that young kindergarten children possess some monitoring skills, but tend to overestimate their performance when asked for ease-of-learning-judgements, and that accurate feeling-of-knowing judgements may be obtained from children as young as 6 years (see Schneider (2008) for a review). A recent study by Balcomb and Gerken (2008) has demonstrated memory-monitoring skills in 3.5-year-old children, in a task originally developed for non-human animals, in which children were given the option to skip uncertain trials on a recognition memory test. Interestingly, young children not only showed evidence for the ability to access their knowledge states in this task, but their memory-monitoring performance also correlated with memory itself.

Young children's knowledge about mental states has also been studied extensively in Theory of Mind (ToM) research (see Flavell (2000) and Kuhn (2000) for discussions of conceptual relations between the two research traditions), and interestingly, the age at which first evidence for metacognition has been found in children is about the age at which children acquire a concept of belief and thus a representational ToM. In a longitudinal study, Lockl and Schneider (2007) empirically investigated links between ToM and metacognition in young children. Both ToM and metamemory development were highly related to language development. However, metamemory at age 5, and mental verb understanding at age 5, were significantly predicted by ToM at age 4 (and to a lesser degree by ToM at age 3), even when language was controlled for. These findings are consistent with

the view that a conceptual understanding of the mental domain, that is, an understanding of the mind as representational, is foundational for both social cognitive and metacognitive development.

A representational understanding of the mind is necessary to understand cases of misrepresentation, such as false beliefs. To understand that an agent has a false belief about a state of reality, a child has to represent the representational relation between the agent and the state of reality. Therefore an understanding of belief requires metarepresentation (Perner 1991). A large body of findings indicates that false belief understanding develops around the age of 4 years and that 3-year-olds' and younger children's failure to solve false belief tasks is due to a genuine conceptual deficit (see Wellman (2002) and Sodian (2005) for reviews). Understanding one's own past false beliefs is as difficult as inferring others' false beliefs (Astington and Gopnik 1991), which supports the view that humans use the same conceptual system, a representational ToM, to represent their own and others' mental states.

Two lines of criticism have recently been advanced against the view that metacognition emerges from a metarepresentational understanding of the mind in the preschool years. First, it has been claimed that metarepresentation develops much earlier than was previously thought, in the second year of life. Second, it has been argued that metacognition should not exclusively be conceived of as metarepresentational, and that the origins of metacognition in infancy may very well be implicit and pre-conceptual. In the following, we will briefly summarize these two lines of argument, and subsequently focus on empirical evidence for metacognition of own ignorance in infancy. Finally, we will speculate about the social construction of metacognition in language-based interaction with siblings and parents.

Metarepresentation in infancy?

Mindreading

Research on the early development of psychological reasoning in infancy has made major progress in the last 15 years (for reviews see Caron 2009; Poulin-Dubois et al. 2009; Baillargeon et al. 2010; Sodian 2011). There is rich and converging evidence for the view that, beginning in the first year of life, infants conceive of human action as goal-directed (Woodward 1998), attributing concrete action goals, as well as dispositions and motivational states to agents, and analysing goals of failed and uncompleted actions (Kuhlmeier et al. 2003; Brandone and Wellman 2009). Most impressively, infants as young as 6 months keep track of what an agent can see when representing the agent's goal, and they do so independently of what they themselves can see (Luo and Johnson 2009). Thus, very early in development, infants integrate a precise representation of an agent's perception with a representation of the agent's goals.

In the second year of life, infants appear to draw inferences from what another person has seen to her subsequent action: 18-month-olds expect a person who has seen the hiding place of a reward to search at the correct location, and a person who was blindfolded to search at the wrong location (Poulin-Dubois et al. 2007). Infants in the second year of life expect another person to act based on her false belief: in violation-of-expectation paradigms infants consistently look longer at an agent who, not having witnessed a transfer of an object from one location to another, searches for the object at its present location, rather than at the location where he last saw it (Onishi and Baillargeon 2005; Surian et al. 2007). Similar findings were obtained when the source of information was not visual perception, but verbal communication or tactile experience, and when the false belief was not a belief about location, but about identity or number (for a review see Baillargeon et al., 2010).

When eye-tracking was used to study anticipatory looking, 18- and 25-month-olds showed evidence for belief-based action anticipations (Southgate et al. 2007; Neumann et al. 2008). Evidence from interactive tasks converges with evidence obtained with looking time methods: Buttelmann et al. (2009) showed that 18-month-old infants interpreted an adult's behaviour differently, depending on whether or not the adult had been present when a toy was transferred from box A to box B. The adult was trying to open box A. If the adult had been absent during the transfer, the infants helped him by retrieving the toy from box B and bringing it to him. In contrast, if the adult had been present during the transfer, they helped him open box A, apparently inferring that, knowing where the toy was located, the adult must have been looking for something else. Thus, there is converging evidence indicating that infants in the second year of life take another person's epistemic state (knowledge, ignorance, false belief) into account when reasoning about her action goals.

Baillargeon et al. (2010) conclude from this body of evidence that infants possess a representational ToM. They argue that infants succeed on spontaneous response tasks (i.e. implicit tasks) that require only the belief representation process, and that they fail on elicited response tasks (tasks requiring an explicit judgement about a false belief task) because of information processing demands of response selection and inhibition associated with such tasks. This rich (metarepresentational) account of infants' mindreading skills is consistent with the idea that infants may be able to represent their own epistemic states in implicit metacognitive tasks. However, to date, much less is known about metacognition than about understanding others' minds in infancy (see later).

Alternative theoretical accounts of ToM in infancy have pointed out that infants' performance on implicit false belief tasks can also be explained in terms of behaviour rules (Perner and Ruffman 2005; Perner 2009). Sodian and Thoermer (2008) and Sodian (2011) have presented evidence for infants' failure to understand the causal relation between perception and epistemic state that is critical to understanding knowledge formation. Rather, infants and toddlers often appear to react to salient situational cues, such as an agent's presence or absence, rather than to epistemically relevant conditions (Dunham et al. 2000). Thus, infants may succeed in some ToM tasks based on automatic reactions to a set of relevant behavioural and situational cues, and fail in situations that require conscious, effortful reasoning. Similarly, Apperly and Butterfill (2009) argue that an early, efficient, but inflexible system of tracking mental states may persist in humans parallel with a later developing, more flexible, and more cognitively demanding ToM.

Self-metarepresentation

Interestingly, in the literature on the development of the self in infancy, it has independently been argued that a capacity for metarepresentation develops around the middle of the second year of life. An important milestone of the early development of the self is mirror self-recognition, assessed with the rouge-test, Amsterdam (1972).While about 40–50% of 18-month-olds recognize themselves in the mirror, almost all 24-month-old children do so (e.g. Asendorpf and Baudonniére 1993; Nielsen and Dissanayake 2004). It has been argued that mirror self-recognition (MSR) evidences self-awareness (Amsterdam 1972; Lewis and Brooks-Gunn 1979; Bischof-Koehler 1991) since MSR demonstrates the child's ability to refer to parts of their body that cannot be seen directly (e.g. the cheek). The child is thereby able to detect the discrepancy between its mental representation and the observed marked mirror image, which can be interpreted as indicating some understanding of how the mirror represents her.

Lewis and Ramsay (2004) consider MSR as a measure of self-metarepresentation. Self-metarepresentation is here used synonymously with the mental state or the idea of 'me'

(Lewis and Ramsay 2004, p. 1821). Michael Lewis (1999, 2003) argues for a distinction between what he refers to as:

> '(. . .) the machinery of myself (the system properties) and the idea of me (a mental state). A young infant can be in a state but may not have an experience of that state. Objective self-awareness—the knower who knows—is not developed until somewhere in the middle of the second year of life, when the self system eventually develops the capacity for metarepresentation. The idea of me then gives rise to mental states in regard to others and in the states of the relation between self and others. This in turn leads to the idea of others also having a "me like me" (. . .)' (Lewis 1999, p. 89).

Contrary to Lewis, Perner (1991) has suggested that MSR requires the ability to form secondary representations which enable the child to simultaneously represent the reality (e.g. the child himself) and the representation of it (e.g. the mirror image). This ability is distinct from the ability to form metarepresentations. According to Perner's theory, infants in their first year of life are limited to generating 'primary representations' of their direct reality. During the second year, children become able to form secondary representations. Secondary representations represent situations decoupled from one's immediate perceptual reality. Multiple models of one situation or event can now be constructed. MSR requires the ability to construct two models, one model of the reality, i.e. the self in front of the mirror, and one of the representation of the reality, i.e. the self reflected in the mirror. It is the understanding of alternative situations that enables children to recognize their mirror image. In the case of the mirror situation, the mirror image is a representation of the real situation in the mirror. Perner (1991) argues that to recognize the correspondence between the real situation and the representation in the mirror, infants need not understand the representational relation between real and represented self. Thus, MSR may be a precursor to metarepresentation, but is not in itself a metarepresentational skill. However, according to Perner's criteria for Mini-Meta-Cognition (see Chapter 6), the ability to entertain alternative mental models of a situation could be seen as a sign of an implicit awareness of the relation between real and represented self.

Non-metarepresentational metacognition

Proust (2003, 2007) argues that metacognition does not require mental state attribution, and is not necessarily metarepresentational. Rather, basic forms of metacognition can be conceptualized as task-specific control/monitoring functions, which require process reflexivity, not self-reflexivity, nor mental-state reflexivity. 'Clearly, a procedural form of metacognition, a "know-how to decide", that is not based on mental concepts and does not need to be made explicit' (Proust 2003, p. 352).

Infants' sensitivity to the degree of learning in habituation paradigms could be accommodated within such a model. Infants' preference for what is novel over what is familiar in habituation or preferential looking paradigms can be interpreted as reflecting a tacit understanding of the degree of learning, such that infants will show a novelty preference at test if they have mastered the information in training, and a familiarity preference if they have not, indicating that infants are not merely associative learners, but that they have some control over how they are learning.

Similarly, infants' engagement in affect-reflective parental mirroring interactions can be seen as involving some degree of monitoring and control over their own and others' cognitive and affective states. It has been claimed that by parental affective mirroring children develop a perceptual sensitivity to internal affect states and an understanding of own states (i.e. the self) as being distinct from the other (Trevarthen 1979). While this intersubjectivist position claims mental state representation in early infancy, Gergely and Watson (1999) proposed that during social contingent interactions the infant senses the causal efficacy of his own actions which is assumed to be a

precondition for later intersubjective understanding that emerges as a result of the maturation of metarepresentational abilities. Within a non-metarepresentational model of metacognition, the infant could well be equipped with some ability to monitor and regulate their own and others' internal states during dyadic interaction, without having to be credited with mental state reflexivity.

It has recently been demonstrated that infants acquire information about the external world from adults' gaze and emotion cues much earlier than was previously thought. Reid and Striano (2007) showed that 4-month-old infants who had watched a video presentation of an adult gazing toward one of two objects, gazed toward the uncued object significantly longer when presented with the same objects at test trials; this novelty preference was also found on the neural processing level. In Reid and Striano's (2007) interpretation, infants benefit from adults' gaze, for instance in reducing the amount of information in the environment, which would be an instance of metacognitive regulation. Such information seeking behaviour clearly becomes intentional towards the end of the first year of life when infants seek emotional information from the adult in social referencing (Campos and Sternberg 1981) or when they *test* for self-other correspondence in a task in which they are confronted with two experimenters reacting contingently to the infants' object-directed actions, one of whom imitates the infant's actions (Agnetta and Rochat, 2004). However, the fact that infants intentionally seek for an emotional or behavioural response when experiencing uncertainty does not imply that they are metacognitively (reflectively) aware of their own uncertainty. Thus, these behaviours can be more parsimoniously explained in terms of a non-metarepresentational self-regulatory mechanism (see Chapters 4, 5, and 6, this volume).

How can such a non-metarepresentational, epistemically implicit control system develop into explicit reflective metacognition? Following Karmiloff-Smith (1992), Proust (2003) hypothesizes a mechanism of representational redescription which eventually results in making knowledge contained in the mind consciously accessible to the mind. Thus, an implicit, non-metarepresentational metacognitive control system is seen as an ontogenetic (and possible phylogenetic) precursor of mentalizing ability. Unfortunately, the developmental relation between early regulatory monitoring and control processes and later explicit metacognition has not yet been studied empirically. Evidence on specific developmental relations would be helpful to counter the argument that Proust's proposal would eventually make just about all behaviour metacognitive (see Perner Chapter 6, this volume).

The early development of metacognition of own ignorance

Early explicit understanding

In experimental studies, an explicit understanding of own knowledge or ignorance, and of the sources of these epistemic states has been demonstrated in children between the ages of 3 and 4 years. Children were asked, for instance, whether they knew what was in a container, when they had or had not been able to look into the container (Wimmer et al. 1988; Pratt and Bryant 1990) or how they found out what was in the container (were they told? Did they see? Did they feel?) (Gopnik and Graf 1988, O'Neill and Gopnik 1991). While language demands may have contributed to the difficulty of some of these tasks, non-verbal tasks were not mastered before the third birthday, either. For instance, Sodian et al. (2006) gave children a choice between a knowledgeable and an ignorant informant (i.e. a person who had seen a hiding event, and a person who had not been able to see it); children younger than 3 years old did not reliably discriminate between the knowledgeable and the ignorant informant, while 2- to 6-year-olds mastered a non-epistemic control task in which they had to choose a person who could help them retrieve an object from a locked box. Similarly, Mascaro and Sperber (2009) did not find evidence for vigilance towards

deception in children younger than 3 years. When a communicator was described as being 'nice' or 'mean', 3-year-olds reliably preferred the benevolent communicator. However, only around the age of 4 years children showed an understanding of the falsity of an utterance by a communicator who was described as a liar.

Evidence for awareness of uncertainty in 2-year-olds

While these tasks required a behavioural choice, that is, an explicit judgement from children, Call and Carpenter (2001) used an indirect or implicit measure of metacognition of own epistemic state by studying children's (and apes') search strategies under different epistemic conditions. In a finding game, stickers were hidden in one of two or three tubes. Subjects saw or did not see the baiting of the tubes; children at the age of 2 years and 5 months looked into the tubes before choosing one more often when they had not seen the baiting than when they had seen it. A rich interpretation of these findings is that children knew when they did not know where the reward was, and acted appropriately to gather relevant information. The authors also discuss the leaner interpretation that children merely knew what they had seen and acted upon this knowledge without having explicitly inferred their own knowledge state. Unfortunately, this line of research into early implicit signs of uncertainty in infancy has not yet been pursued further.

Early mental state language

One reason to believe that 2½-year-old children in Call and Carpenter's (2001) study were aware of their knowledge state is that around the same age first evidence for a nascent explicit knowledge about own knowledge states comes from research on the early use of mental language. In transcripts of natural language, as early as at 18 months of age, children begin to refer to their own mental states and within a 2-month-lag also to the mental states of others (Bartsch and Wellman 1995). While mental state talk mainly consists of emotion and desire terms in the second year of life, cognition terms (e.g. *think*) emerge in the third year, and are used to refer to epistemic states shortly before the third birthday (Wellman and Woolley 1990). When talking about epistemic states such as *know* and *believe*, beginning around 2 years and 8 months, children quite often talk about real-world occurrences, referring to their own mental states about those events and acknowledge that mental states of ignorance and belief can differ from the world. For instance, Adult: 'I thought it was a bus.', Child: 'It's a bus. I thought a taxi.' (Bartsch and Wellman 1995). Or another example, Child: 'I thought it was a crocodile. Now I know it is an alligator.' (Shatz et al. 1983; Papafragou et al. 2007).

In contrast, when making fictional references such as *imagine*, which are not about the factual world at all, children do not show a similar concern about whether the focal mental states are true or not in their fiction-reality contrastives. For instance, Abe (aged 2 years, 11 months): 'I painted on them [his hands];' Adult: 'Why did you?', Abe: 'Because I thought my hands are paper' (Bartsch and Wellman 1995, p. 52). Thus, these utterances demonstrate some understanding of the subjectivity of mental states and of the truth functionality of knowledge and belief.

Evidence for epistemic state representation in joint attention

In a recent longitudinal study, Kristen et al. (2011) found that children's mental state talk at 24 and 36 months of age was predicted by joint attentional skills at 9 and 12 months of age, independently of general language ability. Joint attention has often been interpreted as an implicit Theory of Mind, and recent evidence indicates that, beginning around the age of 12 months, infants do represent others' knowledge states in preverbal communication. Liszkowski, Carpenter, and Tomasello (2007) found that in response to a searching adult, 12-month-olds pointed more

often to an object whose location the adult did not know and thus needed information to find than to an object whose location she knew. While these findings indicate that 1-year-olds represent *others'* ignorance, they also suggest that infants are capable of monitoring their own knowledge in order to non-egocentrically and appropriately communicate with others (see Esken Chapter 8, this volume, for a similar point). In a study by Moll and Tomasello (2004) 12- and 18-month-old infants' search behaviours were studied under conditions of own ignorance. An experimenter gazed behind a barrier at a target outside the infant's line of sight. Some 12-month-olds and the majority of 18-month-olds actively locomoted towards the space behind the barrier to gather information about the object the experimenter referred to, whereas they did not do so in a control condition in which the experimenter looked at an object in the visual space shared with the infant. Thus, by the age of 18 months infants distinguish between conditions of seeing and not seeing in their search behaviours.

Additional evidence for the view that, at 18 months, this distinction is based on some understanding of the epistemic consequences of seeing comes from studies by Butler et al. (2000). These authors presented 14- and 18-months-old infants with an adult experimenter who turned to gaze at a target that was either visible to the experimenter and the infant, or could not be seen by the experimenter because her line of gaze was obstructed by opaque screens, or could be seen by the experimenter because the opaque screens were equipped with transparent windows. The authors found that 14-month-olds tended to turn toward the target above chance in both the no-screen and screen conditions, while 18-month-olds only did so when no screen was present or when the screen was equipped with a window, that is, when the experimenter could see the target. Moreover, about a third of the 18-month-olds leaned forward to look inside the experimenter's side of the screen enclosure in the opaque screen condition. Leaning forward to see what the adult could see may be interpreted to show some understanding of the epistemic nature of seeing, and to suggest that 18-month-olds' discrimination between conditions was not merely due to a cue-based calculation of lines of sight.

The view that infants in the second year of life may have metacognitive experiences of seeing (and not seeing) and possibly also of epistemic states caused by visual perception is supported by findings indicating that infants make inferences from their own visual experience to another person's visual experience. Brooks and Meltzoff (2002) found that 1-year-olds turned less to follow the gaze of an experimenter with closed eyes, but not when she wore a blindfold, suggesting that infants did not know about the effects of a blindfold on visual experience. In a subsequent study, Meltzoff and Brooks (2008) exposed 12-month-olds to a blindfold training in which infants experienced their own gaze as being blocked by the blindfold. This led to a subsequent selectivity in gaze-following (turning less to follow a blindfolded experimenter), compared to control groups who were only exposed to the blindfold (without the experience of their own gaze being blocked by it) or to a trick-blindfold with windows. Even more impressively, 18-month-olds, who without training would not follow a blindfolded experimenter's gaze, after receiving training with a trick-blindfold with windows, showed an increase in turning to look where a blindfolded experimenter looked. Importantly, infants had to use their exclusively first-person experience of not seeing through the blindfold or of seeing through the trick blindfold in order to infer the other person's visual experience, possibly indicating an understanding of the epistemic consequences of seeing in self and others (Meltzoff and Brooks 2008); see, however, Perner (Chapter 6, this volume), for lean interpretations of these findings.

Monitoring the reliability of sources of information

Other people's looking behaviour is an important source of information about the environment for infants. We have argued above that even in the first months of life, infants may use this source

of information to regulate their own information processing. By the second year of life, infants are sensitive to the reliability of a person's looking behaviour and use past reliability as a cue to interpreting future communicative behaviour. Chow et al. (2008) showed that the reliability of a person's past looking behaviour will influence 14-month-olds' decision to follow the person's gaze to a target in front and behind a barrier. First, the infants completed a task in which they watched the experimenter show excitement while looking into a container that had a toy (reliable looker condition) or was empty (unreliable looker condition). Subsequently, they observed the same actor looking at a target object that was visible to the child in front of a barrier (control condition) and at a target object behind a barrier (experimental condition) that was concealed from the child but visible to the actor. Infants in the reliable looker condition were more likely to follow the gaze of the actor to the target behind the barrier than infants in the unreliable looker condition. In contrast, when the target was visible to the infants, there was no difference between the looker groups. In a subsequent study, Poulin-Dubois and Chow (2009) showed that 16-month-old infants respond differently to reliable and unreliable lookers and use this experience to subsequently judge these agents' behaviours in a belief attribution task.

The infants first observed an adult display positive affect (e.g. vocalization, smile) while looking inside a container that contained an attractive object (reliable looker) or was empty (unreliable looker). Although infants from both groups continued to look inside the container, those misled by the unreliable looker became gradually less motivated to verify the contents of the container, as evidenced by their increased latency to open the lid. The infants then watched the same experimenter act as the agent in a non-verbal true belief test modelled after Onishi and Baillargeon (2005). Infants looked longer during the trials in which the adult searched in the wrong place when this person had been a reliable looker in the previous search task. In contrast, the infants who had experienced an unreliable looker looked equally long at the correct and incorrect search in the belief task. The authors conclude from these findings that infants can appraise the reliability of others and encode the identity of an unreliable person. Furthermore, they can generalize their knowledge about a person's unreliable behaviour across different contexts in which the person's gaze is involved. Such generalized expectations that people who had proved to be inaccurate in the past would prove inaccurate in the future had previously only been shown in preschoolers (Harris 2007). These findings indicate that infants use information on the past reliability of a source to regulate their future information seeking behaviours. This may be taken as evidence for procedural (non-metarepresentational) metacognition in the second year of life. It is unclear, however, whether the findings indicate metacognitive awareness of one's own epistemic state in the sense that the reliable looker is conceived of as transmitting knowledge, and the unreliable one as causing a state of misinformation. Lower-level accounts of the learning process occurring during habituation are possible. Recall also that evidence for 'epistemic vigilance' was only found in 3- to 4-year-olds by Mascaro and Sperber (2009). In order to demonstrate epistemic vigilance in infants, we need to show that they can infer the reliability of a source from this person's access to information, and not just from her behaviour (being right or wrong), that is, that they possess some understanding of seeing in others.

A preliminary study of epistemic vigilance in 18-month-olds

In our lab, we have recently conducted a study assessing 18-month-olds' selective use of information from a knowledgeable or ignorant adult (Neumann, 2009). In an interaction task, N = 20, 18-month-old infants interacted with two experimenters, a hider and a cue-giver. The hider sat opposite to the child at a table with two boxes in front of her, while the cue-giver sat in a right angle to both, hider and child. In each trial, the hider placed an opaque barrier between the boxes and the child and then hid a small reward in one of the boxes. Thus, the child could never see in

which box the reward was hidden. The cue giver could observe the hiding events; however, in some trials she attended and in other trials she turned away and demonstrated inattentiveness. Then, the hider removed the screen and the cue-giver directed the child towards one of the boxes. As a dependent variable, we assessed whether infants would follow the cue-giver selectively in trials during which she had access to information about the critical hiding event. Results indicated that infants did not differentiate between valid and invalid cues. In the first half of the experiment (first six trials), infants followed the cue-giver in over 70% of the trials, whether or not she had watched the hiding event. In the second set of six trials infants followed the informant less, again independently of whether or not she had watched the hiding event. This finding is consistent with the interpretation that infants judge persons who are (sometimes or often) wrong as unreliable, but it appears that the reliability judgement was dispositional and not based on an assessment of situational access to information.

Since the task required an explicit response, and since it involved pragmatic constraints, infants' competencies may be underestimated. We therefore conducted an eye-tracking study to test 18-month-olds' selective use of information from a knowledgeable agent. N = 24 infants were familiarized with an animation in which a ghost floated onto the screen and then hid in one of two vertically arranged drawers in a small cabinet. In test trials, the ghost's location was not visible to the infants, but two agents were shown, one of whom watched the hiding event, whereas the other one's visual access was blocked. Subsequently, the two agents simultaneously pointed to the two different drawers (the possible hiding places). Exemplary familiarization and test stimuli are depicted in Fig. 7.1. Results showed that infants spent almost twice as much time looking at the knowledgeable agent's hand than at the ignorant agent's hand, suggesting that they discriminated between the two, and had a tendency to follow the knowledgeable agent's lead. Thus, there is some evidence for an implicit understanding of knowledge formation in 18-month-olds, but no evidence for the ability to access this implicit mental state understanding in a task requiring a behavioural choice, not even when practice and feedback was given. While 18-month-old infants may have had a metacognitive experience of own ignorance in the present task, they seem to have lacked the causal understanding of knowledge formation that is necessary for strategic search for information (selectively follow the person who has seen).

In sum, there is converging evidence from various sources indicating an implicit understanding of seeing and (under some conditions) of the seeing = knowing relation emerging in infants between 12–18 months. This understanding is implicit in communication or in action anticipations and visual preferences, but does not appear to be explicitly accessible in tasks requiring a

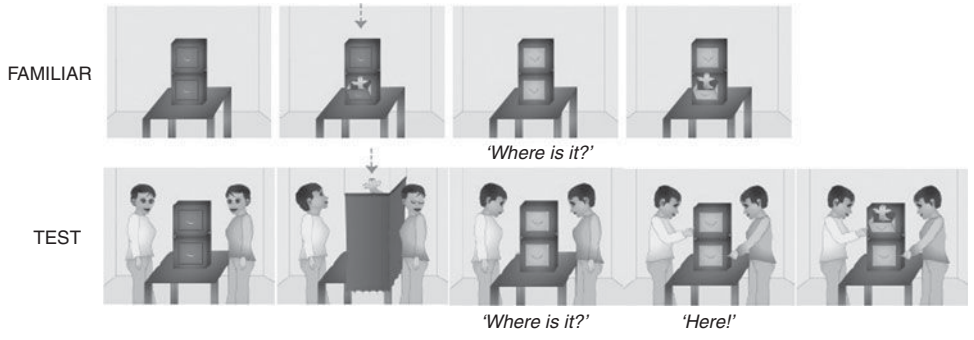


Fig. 7.1 Exemplar stimulus of the eyetracking-task used by Neumann (2009). Reproduced from Neumann, A. (2009). Infants' implicit and explicit knowledge of mental states. Evidence from eye-tracking studies (unpublished dissertation). (See also Colour Plate 4.)

choice or judgement. However, far too little research has directly addressed metacognition of own ignorance in infants and young children.

The role of metacognitive linguistic input in promoting metacognitive development

Metacognitive linguistic input is crucial for children's growing mental state understanding (cf. Dunn et al. 1991). Socioconstructivist theories (Vygotsky 1978) have argued that children may come to internalize metacognitive knowledge by having *collaborative conversations* with knowledgeable others (e.g. Fivush and Baker-Ward 1995; Moore 2006). Specifically, conversations about mental state terms provide children with three kinds of metacognitive input.

First, mental-state discourse includes genuine and clearly articulated utterances in reference to epistemic life, such as *think*, *know*, *believe*, and *wonder* (Bartsch and Wellman 1995). Children are very flexible in their use of these cognitive constructs and early in development (around their third birthday) distinguish between these constructs.

Second, conversations about mental states often focus on explanations of psychological perspectives, especially in relation to psychological causality of behaviour (psychological action explanation) (Wellman 1990). Importantly, it is mothers' elaboration on the mental perspective of others, which is predictive of children's subjective comprehension of the mind (Adrián et al. 2007). For instance, mental state talk with other family members seems to involve indicative, other-oriented mental state reasoning that might help children become aware of and explore the other person's perspective (cf. Brown et al. 1996; Hughes et al. 2010). Illustratively, maternal emphasis on the victim's perspective during conflict situations between siblings has been shown to promote children's understanding of mental states (Ruffman et al. 1999, 2002, 2006). Further, in conflict situations with younger siblings, older siblings seem to refer quite frequently to the inner states of other individuals. Consistently, Jenkins et al. (2003) found that children with an older sibling are advantaged specifically in their talk about epistemic state terms reflecting their comprehension of their own and others' knowledge states.

Third, in older children, mental state language is theoretically related to metacognitive meaning making (Fivush and Baker-Ward 2005). Specifically, the inclusion of emotion and cognitive-processing terms is related to greater context, chronology and theme, as well as overall coherence ratings of children's and adults' narratives, while the inclusion of words like *think*, *know* and *understand* indicate that children have formed organized explanatory accounts of negative events which are integrated with a subjective perspective on their own thoughts and emotional reactions to these events (Fivush and Haden 2005). Sprung (2008) and Sprung and Harris (2010) has shown that children's metacognitive knowledge about mental states and thinking is related to their ability to report on negative intrusive thoughts following a traumatic event. Thus, low levels of talk about mental states may indicate a disruption of the metacognitive meaning-making process. In some children this process is delayed and they can be found to use more internal state language 6 months after experiencing disaster than directly after the event happened (Sales et al. 2005). Further, children coping with stressful medical procedures or emergency room treatment were found to include more internal state language in their narratives of these events (Wolitzky et al. 2005).

In sum, young children refer to the mental, subjective lives of people, not merely their manifest behaviour or physiological properties. Thereby, their utterances are evidential of the fact that they have formed some metacognitive comprehension of their own and others' mental world. They do not equate mental states with observable or objective states, but appreciate their internal, mental quality. Further, they can distinguish mental representations from connections and epistemic and fictional mental states. Finally, when predicting action outcomes based on mental states, they also

understand that the elements of mental state reasoning must be semantically consistent and relevant and not just any desire or belief will lead to any action and reaction.

Conclusions

There is a discrepancy between the burgeoning literature on infants' and young children's ability to read others' minds and the almost complete lack of direct assessments of metacognition in children below the age of 3–4 years. We have argued that some developmental phenomena (e.g. habituation learning, regulatory behaviours), can be interpreted in terms of a model of non-metarepresentational metacognition. Similarly, findings on infants' ability to monitor others' perception and epistemic states may be interpreted as sources of indirect evidence for implicit metacognition, since the tasks require a distinction between one's own and the other's perceptual experience or one's own and the other's knowledge. However, to date, there is no evidence for reflective access to own epistemic states in the second year of life, and little evidence in the third year. Clearest evidence comes from studies of the production of mental state language shortly before the third birthday. Converging evidence on uncertainty monitoring in search behaviours indicates metacognition of own ignorance around the age of 2½ years. Systematic research on the ontogenetic origins of metacognition is needed.

Acknowledgements

The authors would like to thank Frank Esken and Josef Perner for helpful comments on an earlier version of this chapter. The research reported here was funded by a grant from the German Research Council (DFG So 213/27-1,2) and by the ERC project DIVIDNORM.

References

Adrián, J. E., Clemente, R. A., and Villanueva, L. (2007). Mothers' use of cognitive state verbs in picture-book reading and the development of children's understanding of mind: A longitudinal study. *Child Development*, 78, 1052–67.

Agnetta, B. and Rochat, P. (2004). Imitative games by 9-, 14-, and 18-month-old infants. *Infancy*, 6, 1–36.

Amsterdam, B. (1972). Mirror self-image reactions before age two. *Developmental Psychobiology,* 5, 297–305.

Apperly, I. A. and Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116, 953–70.

Asendorpf, J. B. and Baudonnière, P. (1993). Self-Awareness and other- awareness: Mirror self-recognition and synchronic imitation among unfamiliar peers. *Developmental Psychology,* 29, 88–95.

Astington, J. and Gopnik, A. (1991). Understanding desire and intention. In A. Whiten (Ed.) *Natural theories of mind: The evolution, development and simulation of second-order representations*, pp. 39–50. Oxford: Basil Blackwell.

Baillargeon, R., Scott, R. M., and He, Z. (2010). False belief understanding in infants. *Trends in Cognitive Sciences,* 14, 110–18.

Balcomb, F. K. and Gerken, L. A. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Science*, 11, 750–60.

Bartsch, K. and Wellman, H. M. (1995). *Children talk about the mind.* New York: Oxford University Press.

Bischof-Koehler, D. (1991). The development of empathy in infants. In M. E. Lamb and H. Keller (Eds). *Infant Development: Perspectives from German Speaking Countries*, pp. 245–73. Hillsdale, NJ: Lawrence Erlbaum.

Brandone, A. C. and Wellman, H. M. (2009). You can't always get what you want. Infants understand failed goal-directed actions. *Psychological Science*, 20, 85–91.

Brooks, R. and Meltzoff, A. N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, 38(6), 958–66.

Brown, J. R., Donelan-McCall, N., and Dunn, J. (1996). Why talk about mental states: The significance of children's conversations with friends, siblings, and mothers. *Child Development*, 67, 836–49.

Butler, S. C., Caron, A. J., and Brooks, R. (2000). Infant understanding of the referential nature of looking. *Journal of Cognition and Development*, 1(4), 359–77.

Buttelmann, D., Carpenter. M., and Tomasello. M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition,* 112, 337–42.

Call, J. and Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition,* 4, 207–20.

Campos, J. J. and Sternberg, C. R. (1981). Perception appraisal and emotion: The onset of social referencing. In M. E. Lamb and L. R. Sherred (Eds.), *Infant social cognition*, pp. 273–314. Hillsdale, NJ: Erlbaum.

Caron, A. J. (2009). Comprehension of the representational mind in infancy. *Developmental Review*, 29(2), 69–95.

Chow, V., Poulin-Dubois, D., and Lewis, J. (2008). To see or not to see: infants prefer to follow the gaze of a reliable looker. *Developmental Science*, 11(5), 761–70.

Dunham, P., Dunham, F., and O'Keefe, C. (2000). Two-year-olds' sensitivity to a parent's knowledge state: Mind reading or contextual cues? *British Journal of Developmental Psychology,* 18, 519–32.

Dunn, J., Brown, J., and Beardsall, L. (1991). Family talk about feeling states and children's later understanding of others' emotions. *Developmental Psychology*, 27, 448–55.

Fivush, R. and Baker-Ward, L. (2005). The search for meaning: Developmental perspectives on internal state language in autobiographical memory [Introduction to the special issue on internal state language in autobiographical memory]. *Journal of Cognition and Development*, 6, 455–62.

Fivush, R. and Haden, C. A. *(*2005). Parent-child reminiscing and the construction of a subjective self. In B. D. Homer and C. S. Tamis-LeMonda (Eds.) *The development of social cognition and communication*, pp. 315—35. Mahwah, NJ*:* Erlbaum.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. A new area of cognitive-developmental inquiry. *American Psychologist*, 34, 906–11.

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development*, 24, 15–23.

Flavell, J. H., Miller, P. H., and Miller, S. A. (1993). *Cognitive development.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Gergely, G. and Watson, J. S. (1999). Early socio-emotional development: Contingency perception and the social-biofeedback model. In P. Rochat (Ed.) *Early social cognition: Understanding others in the first months of life*, pp. 101–36. Mahwah, NJ: Erlbaum.

Gopnik, A. and Graf, P. (1988). Knowing how you know: Young children's ability to identify and remember the sources of their beliefs. *Child Development*, 59, 1366–72.

Harris, P. L. (2007). Trust. *Developmental Science*, 10, 135–8.

Hughes, C., Marks, A., Ensor, R., and Lecce, S. (2010). A longitudinal study of conflict and inner state talk in children's conversations with mothers and younger siblings. *Social Development*, 19, 822–37.

Jenkins, J. M., Turrell, S., Kogushi, Y., Lollis, S., and Ross, H. A. (2003). A longitudinal investigation of the dynamics of mental state talk in families. *Child Development*, 74, 905–20.

Johnson, C. N. and Wellman, H. M. (1980). Children's developing understanding of mental verbs: Remember, know, and guess. *Child Development,* 51, 1095–102.

Karmiloff-Smith, A. (1992). *Beyond modularity: A* developmental perspective on cognitive science. Cambridge, MA: MIT/Bradford Books.

Kreutzer, M. A., Leonard, C., and Flavell, M. H. (1975). An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development*, 40(1, Serial No. 159).

Kristen, S., Sodian, B., Thoermer, C., and Perst, H. (2011). Joint attention in infancy predicts toddlers' mental state language. *Developmental Psychology*, 47(5), 1207–19.

Kuhlmeier, V., Wynn, K., and Bloom, P. (2003). Attribution of dispositional states by 12-month-old infants. *Psychological Science*, 14, 402–8.

Kuhn, D. (2000). Theory of mind, metacognition, and reasoning: A life-span perspective. In P. Mitchell and K. J. Riggs (Eds.) *Children's reasoning and the mind*, pp. 301–26. Hove: Psychology Press.

Lewis, M. (1999). Social cognition and the self. In P. Rochat (Ed.) *Early social cognition: Understanding others in the first months of life*, pp. 81–98. Mahwah, NJ: Erlbaum.

Lewis, M. (2003). The emergence of consciousness and its role in human development. *Annals of the New York Academy of Sciences,* 1001, 104–33.

Lewis, M. and Brooks-Gunn, J. (1979). *Social cognition and the acquisition of self*. New York: Plenum.

Lewis, M and Ramsay, D. (2004). Development of self-recognition, personal pronoun use, and pretend play during the 2nd year. *Child Development,* 75, 1821–31.

Liszkowski, U., Carpenter, M., and Tomasello, M. (2007). Pointing out new news, old news, and absent referents at 12 months of age, *Developmental Science* 10(2), F1–F7.

Lockl, K. and Schneider, W. (2007). Knowledge about the mind: Links between theory of mind and later metamemory. *Child Development*, 78, 148–67.

Luo, Y. and Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science,* 12, 142–9.

Mascaro, O. and Sperber, D. (2009) The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition,* 112(3), 367–80.

Meltzoff, A. N. and Brooks, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology*, 44(5), 1257–65.

Moll, H. and Tomasello, M. (2004). 12- and 18-month-old infants follow gaze to spaces behind barriers. *Developmental Science*, 7(1), F1–F9.

Moore, C. (2006). *The development of commonsense psychology.* Hove: Psychology Press.

Neumann, A. (2009). *Infants' implicit and explicit knowledge of mental states. Evidence from eye-tracking studies.* (Chapter 7). Ketsch Verlag: Microfiche.

Neumann, A., Thoermer, C., and Sodian, B. (2008). False belief understanding in 18-month-olds' anticipatory looking behavior: An eye-tracking study. *Paper presented at the XXIX International Congress of Psychology*. Berlin: Germany.

Nielsen, M. and Dissanayake, C. (2004). Pretend play, mirror self-recognition and imitation: a longitudinal investigation through the second year. *Infant Behavior & Development,* 27, 342–65.

Onishi, K. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science,* 308, 255–8.

O'Neill, D. and Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology*, 27, 390–7.

Papafragou, A., Cassidy, K., and Gleitman, L. (2007). When we think about thinking: The acquisition of belief verbs. *Cognitio*n, 105, 125–65.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Perner, J. (2009). Who took the cog out of cognitive science? Mentalism in an era of anti-cognitivism. In P. A. Frensch, and R. Schwarzer (Eds.) *Proceedings of the International Congress of Psychology*, 2008, pp. 241–61. Berlin: Psychology Press.

Perner, J. and Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, 308, 214–16.

Poulin-Dubois, D., Brooker, I., and Chow, V. (2009). The Developmental Origins of Naive Psychology in Infancy. *Advances in Child Development and Behavior*, 37, 55–104.

Poulin-Dubois, D. and Chow, V. (2009). The effect of a looker's past reliability on infants' reasoning about beliefs. *Developmental Psychology*, 45(6), 1576–82.

Poulin-Dubois, D., Sodian, B., Metz, U., Tilden, J., and Schoeppner, B. (2007). Out of sight is not out of mind: Developmental changes in infants' understanding of visual perception during the second year. *Journal of Cognition and Development,* 8, 401–21.

Pratt, C. and Bryant, P. (1990). Young children understand that looking leads to knowing (as long as they are looking into a single barrel). *Child Development*, 61, 973–82.

Proust, J. (2003). Does metacognition necessarily involve metarepresentation? *The Behavioral and Brain Sciences,* 26, 352.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 159, 271–95.

Reid, V. M. and Striano, T. (2007). The directed attention model of infant social cognition. *European Journal of Developmental Psychology*, 1, 100–10.

Ruffman, T., Perner, J., and Parkin, L. (1999). How parenting style affects false belief understanding. *Social Development,* 8, 395–411.

Ruffman, T., Slade, L., and Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Development,* 73, 734–51.

Ruffman, T., Slade, L., Devitt, K., and Crowe, E. (2006). What mothers say and what they do: The relation between parenting, theory of mind, language and conflict/cooperation. *British Journal of Developmental Psychology,* 24, 105–24.

Sales, J. M., Fivush, R., Parker, J., and Bahrick, L. (2005). Stressing memory: Long-term relations among children's stress, recall and psychological outcome following hurricane Andrew. *Journal of Cognition and Development,* 6(4), 529–45.

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, 2, 114–21.

Shatz, M., Wellman, H. M., and Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition*, 14, 301–21.

Sodian, B. (2005). Theory of mind. The case for conceptual development. In W. Schneider, R. Schumann-Hengsteler, and B. Sodian (Eds.) *Interrelationships among working memory, theory of mind, and executive functions*, pp. 95–130. Mahwah, NJ: Erlbaum.

Sodian, B. (2011). Theory of mind in infancy. *Child Development Perspectives*, 5, 39–43.

Sodian, B. and Schneider, W. (1990). Children's understanding of cognitive cueing: How to manipulate cues to fool a competitor. *Child Development*, 61, 697–704.

Sodian, B. and Thoermer, C. (2008). Precursors to a Theory of mind in infancy: Perspectives for research on autism. *The Quarterly Journal of Experimental Psychology,* 61, 27–39.

Sodian, B., Thoermer, C., and Dietrich, N. (2006). Two- to four-year-old children's differentiation of knowing and guessing in a non-verbal task. European Journal of *Developmental Psychology*, 3, 222–37.

Southgate, V., Senju, A., and Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science,* 18, 587–92.

Sprung, M. (2008). Unwanted intrusive thoughts and cognitive functioning in kindergarten and young elementary school-age children following Hurricane Katrina. *Journal of Clinical Child and Adolescent Psychology,* 37(3), 575–87.

Sprung, M. and Harris, P. L. (2010). Intrusive thoughts and young children's knowledge about thinking following a natural disaster. *Journal of Child Psychology and Psychiary,* 51(10), 1115–24.

Surian, L., Caldi, S., and Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18, 580–6.

Trevarthen, C. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. M. Bullowa (Ed.) *Before speech: The beginning of interpersonal communication*, pp. 321–47. New York: Cambridge University Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wellman, H. M. (1978). Knowledge of the interaction of memory variables: A developmental study of metamemory. *Developmental Psychology*, 14, 24–9.

Wellman, H. M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Wellman, H. M. (2002). Understanding the psychological world: Developing a theory of mind. In U. Goswami (Ed.) *The Blackwell handbook of childhood cognitive development*, pp. 167–87. Oxford: Blackwell.

Wellman, H. M. and Wooley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, 35, 245–75.

Wimmer, H., Hogrefe, J., and Perner, J. (1988). Children's understanding of informational access as source of knowledge. *Child Development*, 59, 386–97.

Wolitzky, K., Fivush, R., Zimand, E., Hodges, L., Rothbaum, B. O. (2005). Effectiveness of virtual reality distraction during a painful medical procedure in pediatric oncology patients. *Psychology and Health*, 20(6), 817–24.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69, 1–34.

Chapter 8

Early forms of metacognition in human children

Frank Esken

Self-consciousness and metacognition

Currently, there is considerable disagreement about the notion of self-consciousness and especially about the ontogenetic development of self-conscious mental abilities. If, for instance, Gopnik and Meltzoff claim that 3-month-old babies possess genuine forms of self-consciousness (Gopnik and Meltzoff 1994), they may understand the term 'self-consciousness' in a very different way than, for instance, Perner and Dienes, who assume that self-consciousness arises in children between the age of 12–15 months of age (Perner and Dienes 2003). And both conceptions of self-consciousness surely differ from the one which is proposed by Carruthers, who argues not only that children under the age of four do not exhibit any self-conscious mental abilities, but also that they do not possess conscious experiences at all (Carruthers 1989).

The only thing which seems to be more or less uncontroversial about the development of self-consciousness is that at around 4–5 years of age humans develop what is called epistemic self-consciousness, i.e. the ability to entertain higher-order thoughts of the form 'I remember that *p*' or 'He believes that I believe *p*' in an explicit way (i.e. bound to the usage of mental predicates such as 'believe' or 'perceive'). Epistemic self-consciousness is identified with a creature's ability to exhibit *reflexive*, conceptually structured mental states, an ability which is manifested by the ascription of mental states to oneself as well as to others. This sophisticated form of self-consciousness is bound to the ability to grasp what Peter Strawson (1966) calls the objectivity-condition, i.e. to the ability to comprehend the thought that spatiotemporally structured objects exist independently of the experiences that are made with these objects. The development of epistemic self-consciousness in human children is ontogenetically strongly intertwined with the ability to pass standard theory-of-mind tasks (false belief,[1] appearance-reality and representational change). Furthermore, there seems to be credible evidence that this ability is bound to certain linguistic structures (e.g. the use of mental predicates in sentences like 'It *looks* like a stone, but it is a sponge' or 'She *thinks* that the puppet is in the box, but she is wrong'), which are developed only at a rather late stage in human language acquisition (late at the age of 3 or at the age of 4) (cf. e.g. Astington and Olson 1995; Tomasello and Rakoczy 2003).

[1] Some recent studies (e.g. Onishi and Baillargeon 2005) seem to indicate that children possess an *implicit* understanding of false beliefs already on a preverbal level. Because these studies have to use indirect measures such as eye gaze (i.e. infants starting from 13–15 months look longer at scenarios in which a person searches at the correct location despite his false belief) rather than explicit judgements, it is, however, highly debatable whether they really indicate that the infants had to calculate another person's subjective perspective, i.e. that they have to understand something about false beliefs rather than simply to draw first-order inferences and develop expectations from the person's previous to her actual behaviour.

Like 'self-consciousness' the term 'metacognition' is also not without difficulties, because it has been used in two different ways by philosophers and psychologists, respectively:

(A) 'Metacognition' as a synonym for full-fledged mind-reading abilities, i.e. for the ability to attribute mental states to oneself *and* to others in a linguistically structured way. 'Metacognition' in this sense is a metarepresentational ability, i.e. a higher-order thought ability (cf. e.g. Carruthers 1989, 2008; Rosenthal 1990, 1993; Perner 2003; Perner and Dienes 2003), which has to be identified with epistemic self-consciousness. In the following this meaning will be called *declarative metacognition*.

(B) 'Metacognition' as the ability to control and monitor one's own cognitions (cf. e.g. Flavell and Wellman 1977; Koriat 1997, 2000; Proust 2006, 2007).

(B) is used by several authors in a broader sense than (A), since they assume that not all forms of metacognition require a faculty of higher-order thoughts, but that there are metacognitive abilities in humans as well as in some other animals, which do not depend on conceptual, i.e. linguistically structured, but rather on non-conceptual mental contents (cf. Koriat 2000; Proust 2006, 2007). Non-conceptual forms of metacognition are also described as resulting in *epistemic feelings,* such as 'feelings of knowing', 'feelings of uncertainty', and 'feelings of confidence' (cf. Koriat 2000).

A prominent, but not uncontroversial example for such a feeling is the so-called 'tip-of-the-tongue' phenomenon: I know (*by a feeling*) that I ought to know and have known the name of the man I just saw on the other side of the street, but that I will be not able to remember it right now (cf. e.g. Schwartz 1999). The (B) meaning will be called *procedural metacognition* in the following.

An interesting question concerning the two notions of metacognition is how they are related to each other. While it seems to be more or less uncontroversial that declarative metacognitive abilities (A) may be accompanied by, or may result in, certain feelings as, for example, the feeling that one knows the name of the man on the other side of the street, it is much more controversial that there can be creatures who lack declarative metacognition, but who possess procedural metacognition (B).

To illustrate this point, the following example may be helpful:

Surprise as an example for a metacognitive feeling:

Surprise I:

The dog is surprised about the new garden seat in its well-known surroundings and starts barking.

Surprise II:

I am surprised that I only have 5 Euros in my pocket, because I thought that it was much more.

Surprise II surely is a metacognitive ability. It entails a metacognitive judgement ('I thought that I had more than 5 Euros in my pocket') which is connected or leads to a special feeling of being surprised ('I feel surprised'). Surprise I, in contrast, need not be interpreted as involving a metacognitive ability at all, but can be simply based on a perceptual mismatch between a stored representation in the dog's memory and the dog's current perceptual experience of its well-known surroundings which has changed in comparison to the stored representation of it. This mismatch does not mean that there is a mismatch between the dog's *expectation* and his current perception, which the dog *recognizes* as a mismatch. It is just a mismatch between two cognitive states (i.e. a first-order mismatch) which leads to a certain reaction and may involve a certain feeling.

In both cases there may be special feelings involved, but is not certain at all, that in the second case the feeling is also connected to a *metacognitive* ability.

Procedural metacognition in non-human primates?

For a long time experiments on metacognition in non-linguistic animals always had the disadvantage that their results could be explained by means of non-metacognitive, i.e. first-order cognitive abilities.[2]

But the idea of non-conceptual forms of metacognition and epistemic feelings recently gets support from some new results from Kornell et al. (2007). These seem to show, for instance, that rhesus macaques can be trained to make retrospective judgements of their accuracy on perceptual tasks. In one of these experiments rhesus monkeys were shown six pictures, one at a time. Then they were shown nine pictures simultaneously and had to touch the one picture that had been presented previously. After they responded, the monkeys were 'asked' how many tokens they wanted to wager on their response. (The tokens were icons, displayed on the screen, which were automatically exchanged for food rewards). If a subject touched the 'high-risk' icon, they would either gain or lose three tokens, depending on whether their previous response had been right or wrong. If they touched the 'low-risk' icon, they gained one token, regardless of the accuracy of the response just given. The monkeys in this experiment showed a significant tendency to follow correct responses with the choice of the 'high-risk' and wrong responses with the choice of the 'low-risk' icon.

If this experiment shows what it seems to show, some monkeys can be trained to evaluate their remembering abilities, i.e. they can learn to make retrospective judgements of their accuracy on perceptual tasks. These abilities seem to be bound to metacognitive abilities in the (B) sense: in order to perform the task the monkeys have to monitor and to control their own cognitive capacities, i.e. they must be able to reliably express their self-confidence on mastering the task. We normally call cognitive abilities like these (if they are performed by other means, namely by language-based judgements like 'How confident are you, that you got it right?'—'I am quite confident') *metacognitive abilities.* If this interpretation is correct, these monkeys must have learned, in one way or another, to '*reflect*' on their performance abilities, albeit only when highly trained. 'Reflection' in these cases does not mean that they use declarative judgements like 'I know that I saw this picture before', but only that they have learned to register the accuracy of their performances, which engenders *feelings* of confidence or uncertainty. These feelings then would be responsible for their choice of the 'risk' items and exhibit a *procedural* metacognitive ability.[3]

Concerning the topic of this paper the obvious question regarding these findings from primatology should be whether human prelinguistic children are able to pass similar tests as the aforementioned macaques. Unfortunately up to now there is no answer to this and perhaps there will never be an answer, because human children up to the age of 2½ years (an age at which they also have learned basic forms of linguistic communication, which devaluates experiments on non-conceptual forms of metacognition) are much less interested in reinforcement learning than monkeys are, because they are much less interested in being rewarded by food than monkeys are. In other words: they are much less food-junkies than monkeys are.

Early knowledge in infants about mental states/precursors of declarative metacognition

Besides the absence of findings concerning procedural metacognition in prelinguistic children there is accumulative evidence from recent developmental psychology showing that children

[2] Cf. e.g. the experiments on metacognition in monkeys done by Smith et al. (1997).

[3] The 'metacognitive' interpretation of the monkeys' performances in this experiment is nevertheless highly controversial. For alternative, i.e. first-order explanations, see Chapter 5 (Carruthers) and Chapter 6 (Perner). In support of the metacognitive interpretation see Chapter 1 (Couchman et al.).

acquire basic forms of knowledge about some kinds of mental states (i.e. mental states which contain an observational aspect, like visual perceptions, or basic emotions, like fear or anger) long before they acquire theoretical knowledge about these states on a linguistically structured level of explicit knowledge at the age of 4–5 years. Tomasello and Haberl, for instance, found that 12- to 18-month-old infants know which one of three objects a person does not know from past experience (Tomasello and Haberl 2003). Flavell assumes that at this age infants learn that people receive information about the world through vision (Flavell 2004, p. 18). To illustrate these findings, the following experiment might be helpful (Tomasello and Haberl 2003; also compare Moll et al. 2010): 12- and 18-month-olds and an adult played together with two novel objects in turn for 1 min each. Then the adult left the room. While she was gone, the infant and a second adult played with a third novel object. Finally, all three objects were held in front of the infant, at which point the first adult returned and excitedly exclaimed 'Wow! Look! Look at that one' gazing in the direction of all three objects together. She then made an ambiguous request for the infant to hand 'it' to her. Infants of both ages chose the third object—indicating that they knew which of the three objects the adult did not know from past perceptual experience and was therefore requesting from them. However, when the adult had become acquainted with all three objects previously, infants showed no preference for the third toy.

Tomasello and Haberl claim that in order to solve this task, the infants had to understand (1) what the adult knew and did not know in the sense of what she had not become acquainted with from previous experience, and that they had (2) to understand the link between novelty and attention: namely, that people often attend to unknown objects. At first glance this interpretation seems to be unproblematic, but developmental psychologists often go further and state that their findings show that young children between 12 and 18 month of age already develop a first understanding of mental states *as* mental states (i.e. that children pass these tasks because of early forms of metacognitive abilities), which seems to be unwarranted regarding these experiments. To be more precise: a mentalistic reading of these experiments assumes that infants at this age already have learned that perceptions and knowledge are connected with inner experiences, while a behaviouristic reading states that the infants at this age understand them as purely outward behavioural activities. Following the second reading, infants between 12–18 months gradually learn that the behavioural activities of others are connected in a systematic way to gazes and bodily actions, and from this they learn to infer that gazes and actions consist in relations to objects—but they do not yet understand anything about gazes and actions as behavioural activities which are connected with inner experiences. Concerning the much more demanding mentalistic reading, by contrast, one could try to argue that infants between the age of 12–18 months gradually learn to associate, through joint attention interactions, their *own visual experience* of an object and their own bodily interactions with it with the adult's head and eye orientation toward the same object, and come to realize that they are both related to this object via *an inner experience of seeing or knowing it*. Only the mentalistic reading leads to a metacognitive interpretation of these experiments. Due to recent findings that infants are not able to pass so called 'level 1 visual perspective-taking' tasks before the age of 14–16 months,[4] it seems very unlikely that the infants in the experiments just mentioned show an understanding about *mental* states, i.e. that they react not to a mere behavioural but to a mental state. To understand reactions to mental states is more demanding than to understand on a behavioural level what others have and have not seen in 'level 1

[4] In 'level 1 visual perspective-taking' tasks children have to determine what others can and cannot see (e.g. to determine which of two objects an adult was searching for when his view to one of it, but not to the other, is blocked by a barrier). Cf. Flavell 2000, 2004; Moll and Tomasello 2006; Sodian et al. Chapter 7, this volume.

perspective-taking' tasks. Accordingly, the mentalistic reading of these experiments overestimates children's cognitive abilities at this stage of development.

Nevertheless the sketched experiment by Tomasello and Haberl perhaps gives a first interesting hint concerning the development of declarative metacognitive abilities in human infants, especially if it is brought together with the variation by Moll and Tomasello (2007): in contrast to the previous experiment the 14- to 18-month-old infants now observed (1) the adult examine the two known objects individually instead of in joint engagement, or (2) the adult looked on from afar as the infant and an assistant examined the two familiar objects. The adult then left the room while the assistant presented the third object to the infant. In these conditions in which the objects were not shared (i.e. not experienced in joint engagement with the adult), 14-month-olds failed to identify which of the three objects the adult was referring to in her request (cf. Moll and Tomasello 2007; Moll, Carpenter and Tomasello 2010, p. 3). This result is astonishing when compared to the previous experiment which seemed to have shown that even 12-month-old infants know with which objects another person was or was not acquainted with from former experiences. Following these results, infants at 14 months seem to acquire basic knowledge about what others were acquainted with in the past, *only* if this knowledge is acquired in *joint interaction* scenarios (i.e. if the infant is involved in the adult's activities in joint attentional engagement).

Even though the just sketched experiment does not show anything about early forms of metacognition in human infants, it may nevertheless give a first hint concerning the ontogenetic development of these abilities. If infants acquire basic forms of knowledge about what others have and have not seen and were or were not acquainted with in the past, primarily in situations in which they interact with them in joint engagement scenarios, perhaps these findings also have some impact on the development of early forms of declarative metacognition.

On the significance of social interaction for the development of declarative metacognition in human infants

Following developmental psychologists like Philippe Rochat (2001, 2004), the development of metacognitive abilities in humans contains a strongly social component without which this development would hardly be possible. Rochat claims that children at the age of 15–18 months begin to manifest what he calls social co-awareness,[5] and that it is at this age that they begin to develop basic forms of recursive consciousness. Rochat's central claim is the assumption that early forms of metacognition consist in a sense of the self that is exposed to the 'public eye' (Rochat 2001, p. 141).

In a similar way, Tomasello and colleagues emphasize the importance of social interaction for the development of metacognitive abilities in children and claim that what makes humans unique and distinguishes them from other animals is their collaboration activities with others (cf. Tomasello and Rakoczy 2003; Tomasello 2008, chapter 5).

One suggestive piece of evidence for an early form of not yet fully developed metacognitive ability in human children in favour of this assumption comes from Jerome Kagan's experiments on the emergence of signs of anxiety and distress in 12- to 24-month-old children (Kagan 1981, chapter 3): The infants were shown a complex series of steps in the context of an imitative game, which was beyond the range of their behavioural repertoire. Kagan found that at the age of 15 months, but not before, children began to exhibit behavioural signs of distress (like clinging to

[5] With 'the emergence of social co-awareness' in 15- to 20-month-old children, Rochat means that infants learn to understand that they are observed by others (cf. Rochat 2004, p. 141).

the mother, fretting, and crying),[6] which grew quickly with age and reached a peak just before their second birthday. Kagan interprets these results as follows: beginning at the age of 15–16 months children experience an obligation to implement the acts of the model together with an awareness of their inability to perform the action, i.e. they learn to consider and *to reflect* on their performance capabilities by learning what it means to meet rules and instructions (Kagan 1981, p. 50).[7]

Emotions as a case study for metacognitive abilities

While classical approaches to metacognition and self-consciousness more or less neglected that affective states may play any role in the development of these abilities, emotions are rediscovered, as it were, in recent interdisciplinary debates about higher-order mental faculties.

Early infant engagement with others' attention may show, some authors argue, an affective, not yet cognitive and not conceptualized awareness of others as attentive beings, as well as an awareness of oneself as an object of others' attention. Trevarthen, for instance, showed that 4-month-old infants not only respond to attention directed towards them, but also make active attempts to direct others' attention to themselves (Trevarthen and Hubley 1978) and Reddy showed that infants older than 6 months engage with others' attention directed not just at the infant as a whole, but also to specific aspects of their actions (i.e. they are checking on the attention of others after they have completed difficult actions or repeat odd actions that have previously led to laughter, etc.; Reddy 2001). Because the variations of the infants' emotional responses make it rather unlikely that their behaviour can be explained by simple response-reinforcement contingencies, these authors argue that 4- to 6-month-olds exhibit affective forms of metacognition and self-consciousness that are not yet cognitive (i.e. conceptually structured; Reddy 2001, p. 249).

In the opposing corner of the recent debate on emotions there are authors like Lewis, who argue that the earliest self-conscious affects (like embarrassment and empathy) emerge in the middle of the second year of life together with an early, not yet fully conceptual understanding of mental states (Lewis 2001). While Lewis considers primary emotions like anger or fear as non-cognitive, i.e. as stimulus-bound, affective reactions which need not be bound to consciousness at all, he characterizes early forms of secondary emotions like embarrassment as cognitive processes, which require that the creature is able to compare or evaluate its behaviour vis-á-vis some standard, rule, or goal (Lewis 2003, p. 286). Quite similar to Lewis, Rochat claims:

> The expression of embarrassment in front of mirrors by the second year can be interpreted as the first signs of young children's awareness of their public appearance and how others perceive them. [...] By

[6] Distress in this experiment was defined as the occurrence of any one of the following forms of behaviour during the first minute after the model completed her actions: fretting, crying, clinging to the mother, absence of any play with toys during the minute, and protestations indicating the child did not want to play or wanted to leave the room. The most frequent distress reactions were non-verbal and included clinging to the mother, fretting, and crying. The behavioural signs of distress appeared first around 15 months, grew quickly with age, and reached a peak just before the second birthday (Kagan 1981, p. 50).

[7] Concerning Kagan's interpretation that children may become aware of their inability to implement the action, one may object in different ways. (1) The behavior could simply be based on a failure of memory. Kagan rejects this objection, for in many cases in these experiments, when the child had left the mother's side after the distress reaction and begun to play again, he or she displayed an exact or fragmented version of one of the model's prior actions (Kagan 1981, p. 53). (2) A more serious objection may raise the question if the child's awareness of inability really has to be interpreted as an awareness of a lack of *mental* ability, i.e. as a metacognitive competence and not only as an awareness of a lack of *bodily* ability, i.e. as a first-order mental ability.

> showing embarrassment and other so-called secondary emotions, young children demonstrate a propensity towards an evaluation of the self in relation to the social world. (Rochat 2003, p. 723.)

Lewis' and Rochat's considerations fit quite well with Kagan's proposal discussed earlier, namely that early forms of not yet fully developed declarative metacognition involve an evaluative, rule-based component as an early form of recursive consciousness, which entails a strongly socially structured component.

Let us bring Lewis' view on secondary emotions together with Trevarthen's and Reddy's assumption that there are self-conscious and metacognitive affects nearly all the way down in the ontogenetic development of human children and that the classical view on secondary emotions as conceptually structured self-referential mental states only focuses on the tip of the iceberg of these abilities. For this purpose it seems helpful to distinguish between evaluative and non-evaluative forms of affective states like shame, pride and guilt, which are often mixed together. Following this distinction, shame arises as a secondary, conceptually structured emotion in Lewis' sense out of less demanding, non-conceptually structured forms of embarrassment:

1. *Non-evaluative embarrassment (coyness)* Occurs when the child becomes at a very early stage of development affectively, i.e. via a feeling aware of being the centre of attention of others. At this stage the child does not understand on a conceptual level what is going on, but only feels on an affective level quite uncomfortable.
2. *Evaluative embarrassment (shame)* Children become conceptually aware that others are evaluating their actions. This means, that they now understand what it means to evaluate their behaviour vis-á-vis some standard or rule: 'I behaved in way A, but I should have behaved in way B'.

While non-evaluative forms of embarrassment, which also may occur in some non-human animals, like dogs for example, do not possess a metacognitive component at all, evaluative forms of embarrassment imply that the children become aware that others are evaluating their actions and they imply a rule-based component (like 'In this situation I should not have done *a* if I am able to do *b*') as an early form of recursive consciousness. Following this assumption the development of secondary emotions can be described as an external-to internal progression (Lewis 2000) in the spirit of Vygotsky's ideas on the development of higher-order mental functions. The early, purely affective form of embarrassment, i.e. coyness, is driven by environmental-social factors, i.e. being observed by others which results in the uncomfortable feeling of coyness (without any reflection on this situation). The evaluative form of embarrassment, i.e. shame, is driven by the child's reflection on her behaviour in comparison to the requirements which others want it to fulfil.[8] If we follow this view on the development of secondary emotions, evaluative emotions like shame, which children acquire in the middle of the second year, can be regarded as an important step in the development of ontogenetically early forms of not yet fully fledged declarative metacognition in humans: by developing these emotions, human children learn to regulate and control their first-order mental abilities by using basic forms of inner speech (like 'I should have done *b*, but I did *a*'), which they learn by their growing ability to internalize norms and instructions. In the last part of this paper this assumption will be briefly spelled out in more detail.

[8] 'To call a process "external" means to call it "social". Every higher mental function was external because it was social before it became an internal, strictly mental function; it was formerly a social relation of two people. The means of acting on oneself is initially a means of acting on others or a means of action of others on the individual.' (Vygotsky 1978, p. 75.)

Possible relations between the development of executive functions and metacognitive abilities

'Executive functions' in cognitive psychology and related areas are commonly defined as a set of higher-order cognitive processes that modulate and regulate lower-level cognitive (non-mental or first order mental abilities), like motor skills, attention, decision-making, and memory. Up to now, the term is not well defined and the challenges for future research here are very varied. By emphasizing the regulation of information processing necessary to produce higher-order cognitive functions, executive functions are closely related to metacognition, especially to the control aspect of metacognitive processes.[9] Nevertheless not all executive functions are metacognitive in their nature, because metacognition entails a reflexive component concerning lower level cognitive functions. Basic forms of executive functions like for instance hand–eye coordinations or the so-called A-not-B executive task[10] do not entail a reflexive component at all.

Nevertheless there are some theoretical assumptions concerning executive functions which are highly relevant for the research on metacognition. According to the so-called 'Levels of consciousness (LOC) model' (cf. Frye and Palfai 1995; Zelazo and Müller 2002; Frye and Zelazo 2003; Zelazo et al. 2007), which goes back to considerations by Piaget (1937/1952) and Karmiloff-Smith (1992), consciousness can operate at multiple discrete (neuronal) levels, which have a hierarchical structure and develop gradually in ontogeny. Information may be available at one level but not at others. LOC proposes that our common-sense psychology (folk psychology) is the final change in a series of changes in recursive awareness.

According to this model, different levels of consciousness are defined by different control mechanisms (i.e. executive functions) concerning the cognitive control a creature acquires of its sensory interactions (which are, on an elemental level, non-cognitive) with its environment. The LOC-model proposes that human children acquire early forms of recursive consciousness at the end of the first year of life, when they, for instance, learn to search for hidden objects and to point to objects in a proto-declarative way (Frye and Zelazo 2003, p. 254). At this age, infants learn (in a first step) to emancipate themselves from sensorimotorically-driven action schemata and to follow simple rules like 'If I see X, I should do Y'. On a second level of recursive consciousness, which the LOC-model assumes to appear at the age of 18–20 months and which fits well together with our considerations about precursors of declarative metacognition, children learn to understand what it means to meet and to follow rules, i.e. they learn to keep rules in mind and to evaluate their performance in the light of a rule which should be met by their performance (Zelazo et al. 2007, p. 421). Following the LOC-model, children at the age of 18–20 months have developed a concept of the self, but not yet an enduring and objective one. They possess a succession of present-oriented representations of the self, but cannot yet compare previous mental states with current ones (Povinelli 1995, p. 165).

A third important step in the development of full-fledged self-conscious abilities in the LOC-model is assumed to be based on a full fledged metacognitive understanding of what it means to

[9] In fact, some authors, such as, for example Fernandez-Duque et al. (2000), use 'executive functions' as a synonym for 'metacognitive functions'.

[10] In the A-not-B task (cf. Diamond and Gilbert 1989), an experimenter hides a toy under box 'A' within the baby's reach. The baby searches for the toy, looks under box 'A', and finds the toy. This activity is usually repeated several times (always with the researcher hiding the toy under box 'A'). Then, in the critical trial, the experimenter moves the toy under box 'B', also within easy reach of the baby. Babies of 10 months or younger typically make the perseverance error, meaning they look under box 'A' even though they saw the researcher move the toy under box 'B'.

possess a subjective perspective that can differ from the perspective of others. Such an understanding seems to be closely linked to an understanding that one and the same thing can be described in different (and conflicting) ways as it is shown, for instance, by executive tasks like the *dimensional change card sorting task* (DCCS-task): In the DCCS-task children are shown two target cards (such as a red car and a blue flower) and they are asked to sort a series of mismatching test cards, first according to one dimension and then according to the other. For example, they may first be told to sort by colour ('Put the blue ones here and the red ones there') and then be told to switch and, for instance, sort by shape ('Okay, now put the flowers here, put the cars there'). The primary developmental change found in this task occurs between 3 and 4–5 years of age: While 4- to 5-year-olds switch their sorting, 3-year-olds typically continue to sort by the first dimension (regardless of which dimension is presented first). Moreover, they do so even if they are able to *answer questions* about the new rules (cf. Frye et al. 1995; Frye and Zelazo 2003, p. 245).

What is interesting about this case of failing to control one's own actions is that 3-year-olds *know* (on a conceptual/linguistic level) what they should do (they understand the instructions, they can repeat them, and are able to evaluate other children's performances on this task as right or wrong), but they nevertheless do not seem to be able to control their actions with the help of these instructions. More precisely, 3-year-olds are able to form rule pairs for sorting cards by one dimension (one rule, for instance, for sorting by colour and another for sorting by shape). What they do not seem to be able to do, however, is to redescribe objects as being of a different kind; i.e. to understand that one and the same object can be described in different ways. To put the two rules (1. 'Sort by colours'; 2. 'Sort by shape') into effect, the children have to treat an object (e.g. a blue car), either as being described in the one ('as a blue thing') or in another ('as a car') way but not both (cf. Kloo and Perner 2005, p. 54). They fail to redescribe objects as being of a different kind because they fail to *reflect* on conflicting descriptions (i.e. rules) concerning these objects and thus fail to integrate them into a single hierarchical system of rules (cf. Frye et al. 1995; Frye and Zelazo 2003, p. 245).[11]

An interesting point to note about executive-function tasks like the DCCS-tasks is that children are not able to pass such tasks until they reach an age when they also pass standard theory-of-mind tasks. As mentioned earlier, these tasks are assumed to be closely related to the establishing of fully fledged self-consciousness (i.e. epistemic self-consciousness) by philosophers like Strawson and Carruthers, but also by psychologists like Perner and Tomasello (cf. e.g. Strawson 1966; Carruthers 1989, 2008; Perner 1998; Perner and Dienes 2003; Tomasello and Rakoczy 2003; Tomasello 2008). Additionally it is interesting that high-functioning autistic children are able to pass age-related executive tasks which normal children pass between 18–22 months of age, but they fail on more complex tasks like the DCCS-task (Russell 1997). One central assumption proposed by Russell (Russell 1997, p. 287) concerning this finding, is that autistic children, besides their difficulties with affective forms of consciousness, as they are necessary for joint attention activities (cf. e.g. Astington and Olson 1995; Russell 1995 and 1996; Tomasello and Rakoczy 2003), fail to use inner speech to regulate their behaviour. That is to say, they fail to use language as a means of self-monitoring their cognitive capabilities.

[11] 'On this account, a functional process of reflection is required to make a deliberate decision to use the post-switch rules in contradiction to the pre-switch rules, and it is largely through age-related changes in reflection that cognitive development unfolds. Reflection, or reflective abstraction, allows psychological processes at level n to become the contents of level $n + 1$, where they can be integrated with other contents at level $n + 1$.' (Zelazo and Müller 2002, p. 461).

> My claim is, then, that the reason people with autism are challenged by formal tests of executive functioning is that they are unable to use inner speech to regulate their behaviour. I am assuming that, similar to normally developing children in DCCS-like tests (Zelazo et al. 1996), they can represent the rule verbally but their knowledge of that rule does not guide their behaviour. They cannot use language in the service of self-monitoring. (Russell 1997, p. 287.)

Short summary

If the line of thoughts I have tried to sketch in this paper are on the right track, preliminary forms of declarative metacognition in human children develop hand in hand with (1) the growing ability to internalize rules and instructions and (2) with the ability to evaluate one's own performances in the light of these rules coming to the child's mind from outside, i.e. from the human society.

If 2-year-olds begin to play simple rule games and games of pretence such as, for example, 'We play the game *The banana is a telephone*', they play this game by following and internalizing the rule: 'This banana now *counts as* a telephone'. To perform these early forms of pretend play, children do not need fully developed mindreading-abilities of the form 'I know that this is *a*, but I take it now as being *b*'. They do not need an articulated self-concept to be able to engage in this kind of non-deceptive pretend play (compare Brandl Chapter 9, this volume). To keep rules in mind and to evaluate one's own behaviour in the light of rules seems to imply a strongly social component: human infants learn to understand what it means to meet rules by understanding that their actions can be evaluated by others.[12]

If there is something right about this, then early forms of declarative metacognition seems to be quite differently structured than the assumed but controversially discussed procedural metacognitive abilities of monkeys mentioned earlier, which are developed in a solitary, i.e. non-social way. This is not surprising, because 'procedural' and 'declarative metacognition' are used to denote different kinds (also described as implicit versus explicit forms) of metacognition. If besides declarative metacognition there really exists something like procedural metacognition up to now has to be seen as an open question, which further research will have to answer.

Acknowledgements

I am gratefully indebted to Joëlle Proust for giving me the opportunity to work together with her at the Institut Jean-Nicod, Paris, within the EUROCORES Programme 'Consciousness in a Natural and Cultural Context (CNCC)'. I would like to thank Johannes Brandl, Hanjo Glock, Josef Perner, and Beate Sodian very much for helpful comments on this paper.

References

Astington, J. W. and Olson, D. R. (1995). The cognitive revolution in children's understanding of mind. *Human Development*, 38, 179–89.

Carruthers, P. (1989). Brute experience. *Journal of Philosophy*, 86, 258–69.

Carruthers, P. (2008). Metacognition in animals: A sceptical look. *Mind and Language*, 23, 1, 58–89.

Diamond, A. and Gilbert, J. (1989). Development as progressive inhibitory control of action: Retrieval of a contiguous object. *Cognitive Development*, 4, 223–49.

[12] These early forms of evaluation may also involve feelings, but as argued earlier and in contrast to Brandl (Chapter 9, this volume) I see no need to describe them as *metacognitive* feelings, i.e. as feelings that possess a metacognitive dimension by themselves.

Flavell, J. and Wellman, H. (1977). Metamemory. In R. Kail and J. Hagen (Eds.) *Perspectives on the development of memory and cognition*, pp. 3–33. Hillsdale, NJ: Erlbaum.

Flavell, J., Everett, B., Croft, K. (1981). Young children's knowledge about visual perception: Further evidence for the level 1-level 2 distinction. *Developmental Psychology*, 17, 99–103.

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development*, 24, 15–23.

Flavell, J. H. (2004). Development of knowledge about vision. In D. T. Levin (Ed.) *Thinking and seeing*, pp. 13–36. Cambridge, MA: MIT Press.

Frye, D. and Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, 10, 483–527.

Frye, D. and Zelazo, P. D. (2003). Children's action control and awareness. In J. Roessler and N. Eilan (Eds.) *Agency and self-awareness: Issues in philosophy and psychology*, pp. 244–62. Oxford: Oxford University Press.

Gopnik, A. and Meltzoff, A. N. (1994). Minds, bodies and persons: Young children's understanding of the self and others as reflected in imitation and theory of mind research. In T. S. Parker, R. W. Mitchell, and M. L. Brocchia (Eds.) *Self-awareness in animals and humans: Developmental perspectives*, pp. 166–86. Cambridge, MA: Cambridge University Press.

Kagan, J. (1981). *The second year. The emergence of self-consciousness*. Cambridge, MA: Harvard University Press.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Kloo, D. and Perner, J. (2005). Disentangling dimensions in the dimensional change card-sorting task. *Developmental Science*, 8(1), 44–56.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126, 349–70.

Koriat, A. (2000). The feeling of knowing: Some meta-theoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149–71.

Kornell, N., Son, L. K., and Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18, 1, 64–71.

Lewis, M. (2000). The emergence of human emotions. In M. Lewis and J. M. Haviland (Eds.) *Handbook of Emotions*, pp. 304–19. New York: Guilford Press.

Lewis, M. (2001). Origins of the self-conscious child. In W. R. Crozier and L. E. Alden (Eds.) *International handbook of social anxiety: Concepts, research and interventions relating to the self and shyness*, pp. 101–18. John Wiley & Sons.

Lewis, M. (2003). The development of self-consciousness. In J. Roessler and N. Eilan (Eds.) *Agency and self-awareness: Issues in philosophy and psychology*, pp. 275–95. Oxford: Oxford University Press.

Moll, H. and Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, 24, 603–13.

Moll, H. and Tomasello, M. (2007). How 14-and 18-month-olds know what others have experienced. *Developmental Psychology*, 43(2), 309–17.

Moll, H., Carpenter, M., and Tomasello, M. (2010). Social engagement leads 2-year-olds to overestimate others' knowledge. *Infany*, 3, 1–18.

Moore, C. and Corkum, V. (1994). Social understanding at the end of the first year of life. *Developmental Review*, 14, 349–72.

Norman, D. and Shallice, T. (1986). Attention to action. Willed and automatic control of behaviour. In R. Davidson, R. Schwartz, and D. Shapiro (Eds.) *Consciousness and self-regulation: Advances in research and theory*, pp. 1–18. New York: Plenum Press.

Onishi, K. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308, 255–8.

Papineau, D. (2006). Phenomenal and perceptual concepts. In T. Alter and S. Walter (Eds.) *Phenomenal Concepts and Phenomenal Knowledge*, pp. 146–65. Oxford: Oxford University Press.

Perner, J. (1998). The meta-intentional nature of executive functions and theory of mind. In P. Carruthers and J. Boucher (Eds.) *Language and thought: Interdisciplinary themes*, pp. 270–85. Oxford: Oxford University Press.

Perner, J. and Dienes, Z. (2003). Developmental aspects of consciousness: How much theory of mind do you need to be consciously aware? *Consciousness and Cognition*, 12, 63–82.

Piaget, J. (1937/1954). *The construction of reality in the child*. New York: Basic Books. (Original work published 1937.)

Povinelli, D.J. (1995). The unduplicated self. In P. Rochat (Ed.) *The self in infany. Theory and research*, pp. 161–92. New York: Elsevier.

Prinz, J. (2002). *Furnishing the Mind*. Cambridge, MA: MIT Press.

Proust, J. (2006). Rationality and metacognition in non-human animals. In S. Hurley and M. Nudds (Eds.) *Rational Animals?*, pp. 247–74. Oxford: Oxford University Press.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 2, 271–95.

Reddy, V. (2001). Infant clowns: The interpersonal creation of humour in infancy. *Enfance*, 3, 247–56.

Reddy, V. (2003). On being the object of attention: Implications for self-other consciousness. *Trends in Cognitive Sciences*, 7(9), 397–402.

Rochat, P. (2001). The dialogical nature of cognition. *Monographs of the Society for Research in Child Development*, 66(2, Serial No. 265), 133–43.

Rochat, P. (2004). *The infant's world*. Cambridge, MA: Harvard University Press.

Rosenthal, D. (1990). A theory of consciousness. In N. Block, O. Flanagan, and G. Güzeldere (Eds.) *Consciousness – Philosophical and scientific debates*, pp. 729–53. Cambridge, MA: MIT Press.

Rosenthal, D. (1993). Thinking that one thinks. In A. Burri, (Ed.) *Sprache und Denken*, pp. 259–87. Berlin, New York: de Gruyter.

Russell, J. (1996). *Agency: Its role in mental development*. New York: Psychology Press Ltd.

Russell, J. (1997). How executive disorders can bring about an inadequate 'theory of mind'. In J. Russell (Ed.) *Autism as an executive disorder*, pp. 256–304. Oxford: Oxford University Press.

Schwartz, B. L. (1999). Sparkling at the end of the tongue: The etiology of the tip-of-the tongue phenomenology. *Psychonomic Bulletin & Review*, 6(3), 379–93.

Smith, J. D., Shields, W. E., Schull, J., and Washburn, D. A. (1997). The uncertain response in humans and animals. *Cognition*, 62, 75–97.

Strawson, P. F. (1966). *The bounds of sense: An essay on Kant's Critique of Pure Reason.* London: Methuen.

Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: MIT Press.

Tomasello, M. and Haberl, K. (2003). Understanding attention: 12- and 18-month-olds know what's new for other persons. *Developmental Psychology*, 39, 906–12.

Tomasello, M. and Rakoczy, H. (2003). What makes human cognition unique? *Mind & Language*, 18, 121–47.

Trevarthen, C. and Hubley, P. (1978). Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year. In A. Lock (Ed.) *Action, gesture and symbol*, pp. 183–229. London: Academic Press,

Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*, Cambridge, MA: Harvard University Press.

Zelazo, P. D. and Müller, U. (2002). The balance beam in the balance: Reflections on rules, relational complexity, and developmental processes. *Journal of Experimental Child Psychology*, 81, 458–65.

Zelazo, P. D., Frye, D. and Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development*, 11, 37–63.

Zelazo, P. D., Hong Gao, H., and Todd, R. (2007). The development of consciousness. In P. D. Zelazo, M. Moscovitch, and E. Thompson (Eds.) *The Cambridge handbook of consciousness*, pp. 405–32. Cambridge, MA: Cambridge University Press.

Chapter 9

Pretend play in early childhood: the road between mentalism and behaviourism

Johannes L. Brandl

Introduction

The emergence of pretend play in early childhood has been widely recognized as an important source of evidence about children's cognitive development. At the age of 2 years, for instance, children already know that when someone pretends to fill a cup with tea, they can join in the game by pretending to drink from it. How do children get this understanding so early, and what does their knowledge involve? These questions have given rise to a lively debate in developmental psychology. Alan Leslie has argued that understanding an act of pretence, such as 'filling a cup' with imaginary tea, requires a theory of mind and the formation of basic metarepresentational thoughts (Leslie 1987, 1988, 1994). By contrast, advocates of behaviouristic theories have argued that children can follow a simple pretend routine by exploiting only certain behaviouristic clues. From the child's perspective, an act of pretence might be simply a form of behaving *like* a person who fills a cup with real tea, or it may be acting *as if* there were some tea in the pot that children have to imagine. In neither case they would have to know about the mental states underlying such behaviour. Parsimonious explanations along these lines have been advanced by Perner (1991), Lillard (1993, 1994, 1998, 2001), Harris and Kavanaugh (1993), Jarrold et al. (1994), and Nichols and Stich (2000, 2003).

The central claim of this chapter is that we need to look beyond these two alternatives in finding out whether pretend play is an early indicator of metacognitive abilities. I believe we can adequately answer this question only by exploring a third alternative, which steers a middle course between a full-blown mentalistic and a brute behaviouristic account. What makes this third alternative attractive is the suspicion that the other two options are too radical in their claims. They either overestimate or underestimate the cognitive abilities of 2-year-olds. Although no detailed argument for such an intermediary approach has yet been given, one finds several relevant proposals in recent literature on pretend play. Paul Harris has ventured the idea that children's comprehension of pretend play may emerge as a by-product of how children learn to understand goal directed action (Harris 1994). If Harris is right, then a theory of pretend play should look more closely at how children make sense of goal directed action. Josef Perner and Johannes Roessler have followed-up on this idea with a novel explanation of this stage in children's cognitive development. They argue that children can already recognize a reason for acting well before they begin to ascribe beliefs and desires to an agent (Perner 2004; Perner and Roessler 2010). If one applies this claim to pretend play, it suggests that children may somehow grasp the intention that underlies such behaviour, but not necessarily by employing a theory of mind. Hannes Rakoczy and Michael Tomasello have taken this approach, and propose that we look for an explanation of

pretend play that is 'richer than that offered by behaving-as-if theories but not as rich as that offered by Leslie's meta-representational claim' (Rakoczy et al. 2004, p. 397).[1]

How might these proposals add up to a new theory of pretence that avoids the pitfalls of mentalistic and behaviouristic explanations? And what conclusions can we draw from such a theory about the development of metacognition in early childhood? Like all contributors to this volume, I consider the question as to how metacognition develops to be a particularly difficult one. Accordingly, I propose to tackle the issue only after an extensive examination of the difficulties involved in steering a middle course between mentalistic and behaviouristic explanations of pretend play. In the course of that examination, it will emerge that not much attention has been given to the question what role the self-experience of pretending might serve in comprehending acts of pretence. The question as to whether this experience involves a metacognitive feeling is therefore a novel one, and I cannot offer a complete answer to this question here. The small progress I hope to make is to provide a better understanding as to why these issues are both relevant to a theory of pretend play and difficult to resolve.

I begin by developing a basic methodological thesis: young children need not have the same conceptual background as adults in order to have basic competences in producing and responding intelligibly to acts of pretence. In the second section ('Distinguishing trying and pretending'), I turn from methodology to experimental data. I consider evidence provided by Rakoczy and his team that 2-year-olds already distinguish pretend acts from non-pretend acts in which an agent sincerely attempts to achieve a certain goal. In the third ('Leslie's theory of pretence') and fourth sections ('A behaviouristic explanation of early pretend play'), I critically examine the different ways in which mentalistic and behaviouristic theories explain this competence. Finding both kinds of explanation wanting, in the fifth section ('The teleological approach to pretend play') I explore the prospects of a teleological theory of pretence by building on Perner and Roessler's proposal, that young children conceive of intentional actions as goal-directed behaviour. This leads me, in the next section ('The experience of pretending: a metacognitive feeling?'), to consider the self-experience of pretend play as a possible reason why children might find pleasure in the goal of pretending. Finally, I sketch an argument that children's social competence together with this experience makes them aware of their own pretend intentions, and therefore may count as a metacognitive feeling. In this way, pretend play could be an indicator of metacognition even though children do not need a theory of mind to understand it.

The adult's and the child's perspective

Understanding the mind of young children poses a serious methodological problem. Although we often presume to know what children think and feel when we interact with them, minimal theoretical reflection tells us that our adult perspective may be quite misleading in judging how children experience the world. Indeed, children's conceptual knowledge may be quite different from our own; hence we face the constant danger of misinterpreting children's cognitions, by attributing experiences to them that are too sophisticated. This observation is especially critical when we consider children's early pretend play, which begins at 15–18 months (see Piaget 1945/1962). Given that children at this age do not perform well in cognitive tests like the false belief test (see Wimmer and Perner 1983), it would be naive to attribute a cognitively demanding conception of pretend play to them. Some have questioned the reliability of the tests just

[1] The intermediary approach that I am envisaging here is not just a 'mixed' theory that would explain some early components of pretend play in behaviouristic terms, and some later components in mentalistic terms. Nielsen and Dissanayake (2000) have suggested such a mixed approach.

mentioned (see Bloom and German 2000); but that does not affect the point just made. Some reasons to think that children's conception of pretend play is not as elaborate as our own can be given independently of such tests.

There is, first, the obvious fact that our conception of pretence covers not just simple forms of pretend play but also quite complex situations that are beyond what children can comprehend. For instance, when I hold a banana to my ears and start talking, even a 2-year-old child can recognize that I am pretending to make a telephone call. Contrast this with an act of pretence that is much more difficult to decipher. For instance, it may happen that an actor playing Romeo in Shakespeare's play is in fact in love with the actress playing Juliet. How should we then understand his declarations on stage? Does he thereby express his real feelings for the actress or does he imagine being Romeo and being in love with Juliet? Our concept of pretence allows us to ponder such questions, but we surely cannot expect from a 2-year-old child to understand such complex situations in which reality and fiction are blurred.

Secondly, children may not be aware of the fact that pretend play belongs to a family of activities that we call 'pretending' or 'behaving as if' in a broad sense. We say, for instance, that a person *pretends* to be sick when she intends to create the false impression that she *is* (actually) sick, in order to get the day off. But we also say that a hypochondriac pretends to be sick, though she actually believes herself to be sick and does not consciously intend to deceive others. Furthermore, we speak of 'pretence' even when we can attribute neither deceiving intentions, nor false beliefs to the agent. Accordingly, there are at least three cases of 'behaving as if' that we can contrast with pretend play:

1. Cases of intentional deception: a person intentionally generates a false belief in others by pretending something that she knows (or at least believes) to be false.
2. Cases of mistaken behaviour: a person may generate the false impression that it is raining by fetching her umbrella although the purpose of her behaviour is not to deceive anyone.
3. Cases of unintentional deception: animals can deceive others without intending to do so, for instance by creating the false impression of being dead or wounded.[2]

Fig. 9.1 shows how these different forms of 'behaving as if' are systematically related to each other:

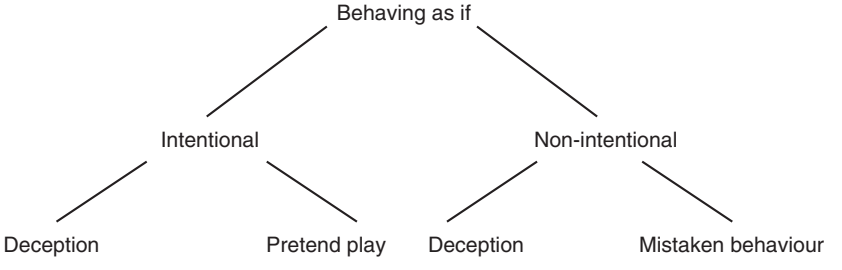


Fig. 9.1 Kinds of 'behaving as if'.

Are children also aware of these distinctions? This depends on what one means by 'being aware of'. Children certainly do not spend much time reflecting on how pretend play differs from other

[2] Whether all deception in animals is unintentional is controversial. When deception becomes flexible and strategic, as for instance in the famous case of the piping plover that feigns having a broken wing to distract predators from attacking its nest, it seems that an intention to deceive is involved (see Ristau 1991).

forms of behaving-as-if. From this we cannot conclude, however, that they are completely unaware of the peculiarities of pretend play. There is a sense in which one might say that when children are pretending, they 'know' that they are neither making mistakes nor engaging in acts of deception. Fittingly, when a child pretends to be sick, the child makes sure that her parents do not actually believe that she is sick. If the child succeeded in deceiving them, the pretend game would end with her being put to bed. The child appears to know this intuitively, and communicates her intention to play a pretend game without thereby exhibiting a clear conception that her intention is non-deceptive.

How children are aware of these distinctions will continue to play a role throughout this paper. It may therefore be helpful to explicitly state at the outset the methodological principles that motivate these further investigations. On the one hand, we must not over-intellectualize children's minds by attributing to them conceptions that they still have to develop. On the other hand, we must not underestimate their cognitive abilities. Children certainly have *some* conception of what it means to pretend playfully; we just do not know yet what this conception may or may not involve.

Distinguishing trying and pretending

Given that children's conception of pretend play is probably simpler than our own, two questions arise. How different is it? And how does their behaviour manifest these differences? The data about children's initial engagements in pretend play largely confirm Piaget's observations. In one famous observation, Piaget reports that his daughter Jacqueline, when she was 15 months old, placed a piece of cloth that vaguely resembled her pillow on the floor. Jacqueline then laid down, put her head on the cloth, repeatedly opened and closed her eyes, all while giving the bystanders a knowing smile. Apparently, Piaget says, Jacqueline was pretending to go to sleep (1945/1962, p. 96).

Most of Piaget's examples concern the production of pretence in early childhood. Recent research, in contrast, has focused on how children begin to understand the pretend acts of others. The data collected in this research suggests that such understanding arises near the end of the second year. Instead of surveying these data here, I focus on a set of experiments that Rakoczy and his colleagues conducted with children between 22–38 months (see Rakoczy et al. 2004; Rakoczy and Tomasello 2006). The aim of these experiments was to find out whether children at this age already grasp the difference between actually trying to do something and merely pretending to do it. That children may find it difficult to recognize this distinction may be seen from the fact that trying and pretending can issue in very similar behaviours. A person can twist a doorknob, for instance, either because she wants to open a door that is firmly locked, or because she pretends to do so. If the person did nothing else but twist the doorknob, how could we be able to tell whether she seriously wanted to open the door or whether she only pretended to do so? It would not be surprising if young children failed to notice that these are different actions. Yet, Rakoczy could show that even the youngest children in this study could solve such a task if relevant clues were provided. Let us see how this worked and what these children achieved.

The experiments used selective imitation as a measure for the range of children's comprehension. They were given demonstrations of actions that they were requested to imitate. While these actions involved the same movements, they differed in gestures and verbal clues that revealed what type of action was performed. In one study, the demonstrator made scribbling movements with a pen that did not produce any graphics because it was covered with a cap. Children could know this from examining the dysfunctional pen beforehand. The demonstrator's action could either be a genuine attempt at drawing something, i.e. it could be a case of acting *as if* one believed

that the pen would work; or it could be an act of pretence, i.e. a case of acting as if the pen worked and as if one did not realize that it was capped. The demonstrator gave the children clues as to which of these two actions he performed by expressing frustration and delight, respectively. When he indicated that he was trying to write something but failed to do so, he showed a furrowed brow and emitted sounds such as, 'Hmm …' and 'What is wrong here?' When he performed an act of pretence, in contrast, he smiled and produced sounds such as, 'Ahh …', while looking at the (non-existing) graphics that he pretended to produce.[3]

The majority of the children tested responded to these demonstrations selectively. After observing an act of sincere effort, most of them imitated the action by also showing signs of frustration when scribbling with the dysfunctional pen. But when an act of pretend play was demonstrated, they imitated this by expressing delight about what they pretended to draw. Although 2-year-olds had a lower success rate than 3-year-old children when imitating pretend play, the differences in their responses were still significant (see Rakoczy et al. 2004, p. 13).

These results are not conclusive, however. For children might mimic the demonstrated behaviour together with the accompanying gestures and emotional expressions, without also grasping their significance. That is to say, children passing the test might simply behave as if they were trying in one case, and pretending in the other, but actually do neither of these things. To rule out such a 'mimicking' explanation, Rakoczy and his team adapted the instructions for children in such a way that they allowed for a greater variation in their responses. In one study, a demonstrator used a closed container as if he wanted to pour liquid into a cup. This movement was again accompanied either with signs of frustration or satisfaction. Children were then allowed to either use the container as it is or to open it with a wrench (see Rakoczy 2004, pp. 389f). When the demonstrator actually attempted to fill the cup, children who opened the container apparently understood his intention because they showed the demonstrator how to do so properly. However, when the demonstrator performed an act of pretence, children refrained from using the wrench and instead imitated the unsuccessful action and responded with appropriate further actions, like pretending to drink from the empty cup. According to Rakoczy, these creative responses are strong evidence that children do not just mimic demonstrations. They must have understood that the demonstrated action was an act of pretence, not of serious trying.[4]

One may still doubt whether this data is conclusive, as have several reviewers of this chapter.[5] Even in these improved experiments, children might merely pick up a certain action scheme that they know how to continue. If they know how to open the container they demonstrate this

[3] It is true, though, that in both cases children observed an act of pretence, since in performing a serious act of trying to write the demonstrator also merely pretended to do so, knowing quite well that the pen did not work. But this fact remained hidden from the children's perspective. They did not know that the demonstrator was specifically instructed to show signs of frustration and therefore had reason to assume that he was really frustrated. From their point of view, it was therefore a case of really trying and not of pretending to try.

[4] Initially, the results seemed to indicate that children younger than three years are not very good in showing these creative responses (see Rakoczy et al. 2004). However, after simplifying the task, even children between 22 and 26 months performed well above chance in this type of experiment (see Rakoczy and Tomasello 2006).

[5] There is, for instance, the possibility that the only difference that children might understand at this age is the difference between 'acting in accordance with the facts' and 'acting contrary to the facts' (see Perner 1995). Since the demonstrator always acted contrary to the facts, it would follow that children reacted only to the different noises that accompanied the movements of scribbling or lifting the teapot. They would miss however the difference between acting on a false belief (unsuccessful trying) and acting contrary to the facts because of pretending that one has a false belief.

knowledge, and if they do not know it they might imitate an act of pretence without comprehending the action that they imitate. That is to say, it still might be true that children react differently to what for them is just a pattern of movements that they replicate. Such alternative interpretations are difficult to rule out completely. Nonetheless, I maintain that children's flexible responses provide good reason—even if not conclusive evidence—that they recognized a difference in the demonstrated actions. I grant this here because I will later argue that children at this age employ a teleological form of reasoning which gives them the capacity to do this. First, however, I need to prepare the ground for making this claim by considering how the two leading theories of pretend play would explain the data at hand.

Leslie's theory of pretence

Given that 2-year-old children already seem to know the difference between trying and pretending, psychologists face the following puzzle. Theorists have widely assumed that children at this age do not yet understand that people's actions depend on how they represent things to be and that children therefore do not yet pass a standard false belief test and are also incapable of intentional deception (see Perner 1991). But isn't the same ability, namely to understand how people represent the world, also needed for grasping an unsuccessful act of trying when the agent holds a false belief, as well as for grasping an act of pretence when he knows the truth and still acts as if he were ignorant of it? How could 2-year-old children distinguish between trying and pretending if their cognitive abilities are still too limited to understand the psychological underpinnings of behaviour based on false beliefs or deceptive intentions?

Leslie's theory of pretence solves this puzzle with a radical move. He challenges the standard view by postulating a basic 'theory of mind mechanism' (ToMM for short) that is in place well before the age at which children have been assumed to become 'mindreaders' (see Leslie 1994). Equipped with this mechanism, even 2-year-olds already form beliefs about what others believe, and also beliefs about what others merely pretend to be the case. It therefore cannot be the lack of these metarepresentational capacities that explains why children up to 4 years tend to fail a false belief test and why they have difficulties in deceiving others (see German and Leslie 2001). In support of this claim Leslie points out that children already have to solve a representational problem when they begin to understand pretend play. They need to avoid a conceptual confusion due to 'representational abuse', as Leslie calls it (see Leslie 1987, p. 415; 1988, p. 22). When they see their mother pretending to make a telephone call with a banana, for instance, they need to understand that the banana is not a new kind of telephone that they encounter for the first time. The only way to avoid such confusion, according to Leslie, is to realize that mother herself does not believe that the banana is a telephone. Without a ToMM, children would revise their concept of a telephone in confronting such pretend behaviour. This is Leslie's central thesis: 'Being able to pretend and to understand pretence in others requires mastery of exactly the same 'logical' structures as understanding mental states. One could say that early pretend play is a primitive manifestation of a theory of mind' (Leslie 1988, p. 24).[6]

From this point of view the data reported by Rakoczy et al. are not surprising, as they must appear to be to proponents of the standard view. On the contrary, they confirm that children have an early psychological competence that they rely on when imitating an act of trying in one way,

[6] The conclusion that Leslie draws here may be unwarranted, as critics have pointed out. One can find the logical peculiarity of mental state ascriptions to which Leslie refers here also in other forms of discourse, e.g. in describing past and future events. Understanding such descriptions does not count as a primitive manifestation of a theory of mind, however (see, e.g. Perner 1991, pp. 62f).

and an act of pretending in a different way. They pass this test because their cognitive systems are complex enough to generate the following representation of the different actions involved:

Trying: the demonstrator acts as if he believes that the pen is working normally.

Pretending: the demonstrator acts as if the pen were working normally although he does not believe it.

How much support is there for Leslie's view? If his theory is correct, 2-year-old children already operate with concepts that allow them to grasp the *mental* difference between acting on a false belief and acting as if one had a false belief. There would be no essential difference between how they analyse such situations and how we do from our adult perspective since they would also do it in mentalistic terms. Two objections can be made here that raise considerable doubts that this is true.

First, Leslie does not deny that children have beliefs and desires long before they begin to understand what these mental states are (see Friedman and Leslie 2007, pp. 107f). But he does deny that such an asymmetry between having and understanding exists in the case of pretending. Pretending is special in this respect, as he says: it is the 'most striking fact about the development of pretence that [...] when the child acquires the ability to pretend herself she simultaneously acquires the ability to understand pretence in others' (Leslie 1988, p. 29). While in general being in a mental state does not imply the ability to attribute such states to others, in the case of pretending these two conditions coincide according to Leslie: 'understanding pretence in others is simply part and parcel of being able to pretend oneself' (Leslie 1987, p. 416). However, the empirical evidence does not support this exceptional claim about the special nature of pretend play. Children seem to discover the joy of pretending several months earlier before they manifest an understanding of the pretend nature of such acts when performed by others (see Harris and Kavanaugh 1993, pp. 75f).

In addition to this empirical objection, it is also unclear that Leslie has a successful argument for his claim. The basic structure of his argument seems to be the following:

First premise: one cannot pretend that *p* without knowing that one pretends that *p,* which means that one must truly believe, 'I herewith pretend that *p*'.

Second premise: one cannot form the belief 'I pretend that *p*', unless one can also form similar beliefs of the form, 'X pretends that *p*', about other agents.

Conclusion: one cannot pretend that *p* without understanding acts of pretence performed by others.

This argument is far from persuasive. Doubts arise here concerning both the truth of the premises and the legitimacy of the inference from those premises. Even if one grants the initial assumption that a pretender has to know about his pretence, one can reject the implication that his knowledge requires an explicit belief of the form, 'I herewith pretend that *p*'. The second premise remains doubtful because a special mechanism may enable the formation of first-personal beliefs without having the ability to form similar beliefs about other agents. So there are neither empirical nor sufficient theoretical reasons for claiming such a close tie between pretending and understanding pretence. This pillar of Leslie's theory is extremely shaky.

It may seem unfair, however, to criticize Leslie's theory while leaving out most of its complexity. Let me therefore add a further consideration that casts doubt on the notion that further complexity might save the theory. As we have seen, Leslie explains children's competence in pretend play with a ToMM in virtue of which children may form metarepresentational thoughts. Leslie is careful to point out that this does not entail that children have a 'representational theory of mind', as this term is often understood (see Leslie 1994, p. 217). To mark this difference, Leslie introduces

the term 'M-representations' for a kind of metarepresentations 'light,' so to speak, which enable children to reason about belief states without employing notions like 'reference' and 'truth' that belong to a fully developed representational theory of mind (see Leslie 1988, pp. 31ff; Friedman and Leslie 2007, fn. 3).

Whether young children exemplify such a weaker form of metacognition is an interesting point to which I will return later (see section 'The experience of pretending: a metacognitive feeling?'). The important point at present is whether the possibility of such weak metacognition makes Leslie's theory more plausible, especially in light of empirical data that seems to undermine it. I do not believe that it does; and we can see this via comparison with a similar proposal by Perner, Baker, and Hutton. These authors argue that as long as children lack a full-blown representational theory of mind, their concept of belief will be immature and not enable them to clearly distinguish between believing and pretending that something is the case. One might attribute them merely what these authors call a concept of 'prelief' (see Perner et al. 1994). In parallel to this proposal, Leslie's theory could be interpreted in such a way that cognitive systems that operate with M-representations only have mental concepts that are as non-discriminating as the concept of 'prelief'. But if this is so, then our puzzle arises again: how could children distinguish trying from pretending with such immature concepts? When they observe what the demonstrator does, they may form only the following thought: he is acting as if the pen were working because he *preliefs* that it is working. This description fits the case of trying as well as the case of pretending, and so children's different reactions in each case would again remain unexplained.

At the beginning of this section, I intimated that Leslie's theory seems to be perfectly suited for explaining Rakoczy's data. This claim has now turned out to be too optimistic. Once it is granted that children's mental concepts might be quite different from our own, the alleged explanatory power of the theory vanishes. So one might again question the validity of the data instead of trying to explain them. That is not the path I want to take, however, for I will argue that there is a plausible teleological explanation of these data. To motivate this alternative view, however, we must first consider the primary opposition to Leslie's theory—a behaviouristic theory of pretence—to see what explanation such a theory has to offer.

A behaviouristic explanation of early pretend play

The term 'behaviouristic', as used in this context, is rather misleading, for one might think that so-called behaviouristic theories attempt to explain pretence without making reference to any mental states whatsoever. That is not the case. Rather, what these theories claim is merely that *children* need not employ concepts for mental states when they pretend. Behaviouristic theories also claim that children can make sense of others' pretend behaviour without invoking mental concepts. The behaviourist claims that children themselves take a behaviouristic stance, not that full-stop behaviourism is the stance that *we* should take in explaining children's cognitive and metacognitive abilities.

One way to interpret this claim would be to say that children have a basic competence in participating in such games, but do not conceptualize the playful actions from the external perspective of an observer.[7] For instance, a child may show its competence by telling a newcomer who enters the room: 'You have to be quiet because Mum is making a telephone call!' If she does this with a knowing smile while Mum is holding a banana to her ear (recall Jacqueline's smile in Piaget's earlier example), the child signals that she understands her mother's action to be pretended. Taken literally, the child's statement would be false, since it does not mention that it is a

[7] Thanks to Johannes Roessler for making me aware of this possible interpretation.

pretend action. No misunderstanding occurs, however, as long as the newcomer takes it to be a statement from the internal perspective of a participant.

Though plausible, this interpretation does not fully capture what behaviouristic theories claim. Those theories do not deny that children can understand pretend play from the external perspective of an observer who explains the game to others but does not participate in it. So what is their claim? Let us return to Leslie's problem of representational abuse—that is, the problem of leaving one's concepts unchanged in pretend situations despite applying them in inappropriate conditions. Advocates of behaviouristic theories are also concerned with this problem, but they do not take it to warrant postulating a 'theory of mind module' that helps children to solve this problem. Instead, they claim that they can solve this problem with a simpler functional device, as we shall see, that does not imply possession of mental concepts. It is this difference that sets their view apart from a mentalistic theory of pretence. In what follows, I will use the functionalist architecture proposed by Shaun Nichols and Stephen Stich to explain the motivations for and consequences of this difference.

A key element in the model used by Nichols and Stich is the so-called 'possible world box' (PW-box for short) that allows a cognitive system to 'quarantine' propositions (see Nichols and Stich 2000, 2003). A PW-box contains propositions that a cognitive agent can use for reasoning without also having to believe that these propositions are true. Nichols and Stich argue that once this mechanism develops, children are cognitively prepared for pretend play. A child can now put the proposition 'the banana is a telephone' into her PW-box, use it to conclude that Mum is making a telephone call, and act in accordance with this conclusion without believing the proposition to be true.[8] In keeping her imaginations separate from her beliefs, the child effectively avoids commitment to the mistaken belief that bananas are both fruits and telephones.

How could such a fully cognitivist model qualify as a behaviouristic explanation of pretend play? The crucial difference between this model and Leslie's model concerns the retrieval process of propositions. Nichols and Stich assume that it is possible for a child to retrieve a proposition '*p*' from her belief box—in other words, activate her belief that *p*—without forming the complex thought, 'I believe that *p*'. Equally a child can retrieve a proposition from her PW-box without forming the complex thought, 'I am imagining that *p*', or 'I am pretending that *p*' (see Nichols and Stich 2003, pp. 50ff.). In making the retrieval of propositions generally as simple as that, Nichols and Stich undercut Leslie's claim that in order to understand pretend play one must already possess a basic theory of mind capacity. The PW-box accomplishes this just as well, without requiring children to have a ToMM.

Let me return now to the question as to why this model does not restrict children to a merely internal perspective in comprehending pretend play. How could a child equipped only with a PW-box make sense of pretend play by others? Doesn't that require that children make this extra step and think that the other person is pretending that *p*? Advocates of behaviouristic theories do not deny that such an extra step is required, but they would claim in the present example that the content of such a thought is as follows: 'Mum is behaving as if the banana were a telephone,' or more explicitly, 'Mum is behaving in a way that would be appropriate if the banana were a telephone' (see Nichols and Stich 2003, p. 53).

Having explained the key idea of a behaviouristic theory of pretence, we can now turn to its evaluation. Leslie thinks that any theory of this kind is hopeless because it generates too many

[8] Another way to describe the special nature of this reasoning process would be to allow that certain concepts are used in a non-literal way. A child could then conclude from the fact that the banana is 'a telephone' that Mum is making 'a call'. In this case, a mechanism would be needed to 'quarantine' the literal meaning of these concepts while still allowing such inferences.

pretend descriptions. He calls this *the problem of overextension* (see Friedman and Leslie 2007). In its strongest form, the objection claims that if children conceive of pretence from a behaviouristic point of view, they could interpret *any* action as an act of pretence. No such tendency has been observed, however, and so the theory is empirically false. I think that, at least in this strong form, the objection goes too far. The best way to show this is to make clear that a behaviouristic theory suffices to explain how children distinguish trying from pretending. Responding to the objection in this way, however, does not absolve behaviouristic theories of more significant difficulties. But in order to uncover such difficulties, it will help first to show why Rakoczy's data do not pose a serious problem for these theories.

Consider again the example in which children had to recognize whether the demonstrator was trying to make a drawing or merely pretending to do so. If children could observe only the demonstrator's movements—i.e. scribbling with a pen—they would have no way to determine the intent of his demonstration. Once the demonstrator provides them with additional information however, namely his frustration or delight, children can integrate this information into their behaviouristic reasoning thus:

> When the demonstrator shows signs of frustration, he behaves in a way that would be appropriate if the pen were not working. But when he shows signs of delight, he behaves in a way that would be appropriate if the pen were working.

This example shows the way in which children have the ability to distinguish between acts of trying from acts of pretending without having to make any reference to the mental states of the pretender. Leslie's objection that any action could be described as an act of pretence from a behaviouristic point of view does not hold for actions that are recognized as serious attempts to do something. But the integration of additional information will not work for the behaviourist in all cases. It breaks down when the clues given are sufficiently ambiguous. Such cases are both easy to imagine and quite common. Suppose there is a basket that looks as if it were filled with fruits, but contains only painted pieces of wood that look like fruits. Children may know this while they observe an actor who is ignorant of this fact. Again, we can imagine that two actions are demonstrated. The agent either packs these 'fruits' into his lunch box, smiling because he is looking forward to eating them later; or he is packing the fruits with a 'knowing smile', indicating that he knows that they are not real. In this case, the correct interpretation of the smile is the only way to find out which of the two actions has been performed. But no such difference in interpretation is available from a behaviouristic point of view, since it is equally appropriate to smile in both cases. Therefore, a behaviouristic theory has to predict that children at the age of 2 are not yet in a position to differentiate between pretending and acting on a false belief in such cases. For them, both actions would be cases of pretending in the sense of acting-as-if, with no possibility of differentiating between these two cases.

We do not know from existing data whether this prediction holds. It may be that children need something like a theory of mind for distinguishing a smile indicating a joyful expectation from a knowing smile indicating an act of pretence. It seems to be a weakness of behaviouristic theories, however, that they leave this as the only possible explanation. There is no reason to think that when children get the idea expressed in a knowing smile, they suddenly switch from a behaviouristic to a mentalist understanding. Accordingly, the main objection I have against such theories is that they tend to equate 'pre-mentalistic' with 'purely behaviouristic'. That equation does not do justice to the gradual progress observed in children's cognitive development. I take it that a similar dissatisfaction with behaviouristic theories leads Rakoczy and Tomasello to suggest that 2-year-old children might 'grasp the intentional structure of pretending as a specific non-serious action form, different from other forms of behaving-as-if.' (Rakoczy and Tomasello 2006, p. 558).

At the same time, however, Rakoczy and Tomasello do not find reason to adopt a mentalistic theory either.[9] If the evidence shows that even 2-year-old children understand the smile of a pretending person as a knowing smile, some further explanation is called for.

The teleological approach to pretend play

It is a truism of folk-psychology that persons engage in actions with the intention to achieve certain goals. The problem with such truisms is that one can interpret them in incommensurable ways. The conception of intentional action standardly employed in the philosophy of action rests on the assumption that intentions are grounded in subjective reasons. Unfortunately, this prevalent interpretation of intentional action has occluded the fact that we can understand intentional action also as behaviour guided by objective reasons only. The omission here is deplorable, as Perner and Roessler have pointed out, because it may be precisely in this way that children comprehend intentional actions. If a teleological interpretation of common sense psychology also leads to a better explanation of the emergence of early pretend play, as I will argue, this makes Perner's and Roessler's claim even stronger.

First, however, we must clarify the different interpretations of intentional action that are at work here. The core thesis of the standard view says that intentions are subject-relative states grounded in beliefs and desires. One intends to do *H* if one believes that doing *H* will have certain desirable effects. Take, for instance, a soccer player preparing to shoot a penalty kick. The soccer player believes that by shooting a goal, he can contribute to winning the game for his team. That is one way to understand that he acts intentionally: one can 'derive' his intention from a belief-desire pair: a desire to win, coupled with a suitable belief as to how to satisfy this desire.

Soccer players and soccer fans do not have to be so elaborate in their reasoning, however. They can understand this situation simply on the grounds that they know the purpose of the game and the rules for performing a penalty kick. Using this knowledge they can make sense of the player's behaviour in terms of what Perner and Roessler call 'objective reasons', instead of appealing to subject-relative beliefs and desires that tend to vary from person to person. The term 'objective', in this context, carries no metaphysically problematic assumption about the objectivity of values. It merely captures the idea that within a given community certain values are shared and experienced as part of social reality. It is precisely in this specific sense of 'objective' that soccer fans share the knowledge that scoring a goal is something objectively good. Of course, wanting to win the game is an internal mental state, and is in this sense subject-dependent. Yet it can be *experienced* as an objective value shared by everyone participating in the game. In the same way, we can distinguish between instrumental beliefs of individual subjects and instrumental knowledge that is taken to be common ground. Each individual player may hold the belief that scoring leads to winning the game. What matters, however, is that they together take the rules of the game to be mutually known facts. Hence there is no need to incorporate a subject-relative notion of belief into one's explanation. Common knowledge suffices to explain the soccer player's preparing for a penalty kick.

There are many questions here that would need a more detailed discussion. How do agents come to share certain values? How do these common values become goals for them to pursue

[9] In commenting on this proposal, Friedman and Leslie simply deny that there is a further option to be considered. They take any theory that is not mentalistic to be a purely internal explanation of pretend play that is too weak to explain the data. In their view, the approach suggested by Rakoczy and Tomasello does not gain anything because it is still a basically behaviouristic theory that falls victim to the problem of overextension (Friedman and Leslie 2007, pp.119f).

individually? How do they pursue these goals on the basis of a common knowledge about how these goals may be realized? And how does this all work without bringing in subjective beliefs and subjective preferences? Without entering this discussion here, I can merely sketch the basic principle of a teleological account of intentional action that I will then also apply to pretend play. The schema that governs our practical reasoning from a teleological point of view can be stated as follows:

1. It is a good thing to bring it about that *p* (e.g. winning a game).
2. By doing *H*, one can bring it about that *p* (e.g. by making a penalty kick).

Therefore: if an agent is in a position to do *H*, he will do it.

What differentiates this principle of practical reasoning from the standard conception of intentional action is this: on the standard view one would interpret premises (1) and (2) as specifying the content of an agent's desire and instrumental belief, because an agent who does not hold these desires and beliefs would have no reason to do *H*. On the teleological account, by contrast, one can make such a prediction by taking the premises as stating two objective facts. Of course, if an agent were completely unaware of these facts, i.e. if he did not know that it is a good thing to bring it about that *p* and that doing *H* will have this effect, he would not have a reason to perform this action on this account either. A teleological explanation, too, must therefore assume that agents are aware of these facts in some way. But it need not cash out this assumption in terms of beliefs and desires that vary from agent to agent. A connection between these objective facts and a particular agent who takes them as his reasons for action can also be made in terms of 'what everyone wants' and 'what everyone knows.'

That we find in our folk psychology such a teleological conception of intentional action is highly significant for understanding children's cognitive development, as Perner and Roessler as well as other developmental psychologists have pointed out (see Csibra and Gergely 1998; Gergely and Csibra 2003). Children may acquire that part of our folk psychology quite early. They will then be aware that agents have reasons to do certain things not because they have subjective beliefs and preferences, but because they engage in an objectively reasonable form of goal directed behaviour. In the words of Perner and Roessler: 'children find actions intelligible in terms of fully objective reasons, relativized neither to the agent's instrumental beliefs nor to her pro-attitudes' (Perner and Roessler 2010, p. 205). The significance of this claim becomes clear when one considers the classical false belief task in which 3-year-old children give incorrect, but perfectly reasonable answers. The task in this test is to figure out where a character named Maxi will look for his chocolate after it has been transferred from one location (drawer) to another location (cupboard), without Maxi being able to observe this transfer. Three-year-old children tend to say that Maxi will search for his chocolate at the location where it is, instead of making the correct prediction that he will go to where he still believes it to be. From a teleological point of view, this incorrect answer is not at all irrational because children reason as follows: 'Maxi needs his chocolate. (Or: it is important, or desirable, that Maxi obtains his chocolate.) The way to get it is to look in the blue cupboard. So he should look in the blue cupboard.' (ibid.)[10] Children fail the test because they lack the concept of a false belief, not because they do not see any rationality in the action that they predict. While the first conclusion is the one that gave the false belief test its name, the second

[10] On the interpretation intended here, 'he should look' can be replaced by 'he intends to look'. The notion is meant to specify what Maxi should do from the objective point of view that children take for granted before they consider the possibility that Maxi's thinking may be different because he lacks relevant information. Only if this is taken into account, can it happen that Maxi intends to do something that differs from what he should do.

conclusion is no less important: the test also shows that children up to a certain age construe intentional action in terms of objective goals and in terms of common knowledge about how to reach these goals.

Let us now consider whether we may usefully employ this teleological kind of explanation to illuminate the phenomenon of pretend play. Can we account for children's early competence in such play via a reasoning process that accords with the teleological principle stated earlier? In answering this question, I will proceed in two steps. First, I consider how a teleological reasoning process could enable children to *form* the intention of pretending to do something. Secondly, I consider how they might *understand* an act of playful pretence when performed by others.

It will be important to keep in mind that the term 'forming an intention' need not refer to a self-reflective mental activity. If this were so, we would have to assume that children begin to engage in pretend play without forming an intention to do so. It would be strange to say, however, that such actions are not intentional at all. Therefore we should allow the term 'forming an intention' to have a broader meaning that may or may not include the reflective awareness of having a plan how to act. Following the soccer example above, we might say that children can 'get ready' to perform an act of pretence once they know the purpose of the game and how it should be played in a given context. It is not as simple as that however, since in the case of pretend play we cannot refer to a set of rules that are constitutive of such games and that children learn. If one applies the teleological principle of reasoning here in a simple-minded manner, pretence would always lead to frustration. Consider again a child that picks up an empty teapot in order to pretend filling a cup with tea, and suppose that the child thereby reasons as follows:

1. It would be a good thing if this cup were filled with tea.
2. Holding the teapot over the cup will bring it about that the cup gets filled.

Therefore: hold the teapot over the cup.

If this were how children form an intention—i.e. get ready—to pretend, then that would imply that children are in an awkward situation when they pretend: for their ways of forming intentions would seem to imply that the child will not reach its goal of filling the cup with tea, given that there is no tea in the pot. Why then should any child find pretend play enjoyable? Notice that this problem would also arise if one were to interpret premises (1) and (2) as specifying the content of a subjective belief and desire. Clearly, we need some other conception of what the goal in pretend play is.

The answer I want to suggest in line with the teleological conception of intentional action is that children conceive of pretending as an objective value. That is to say, they conceive it to be a good thing to perform an action even if the conditions for doing it successfully are not, or cannot be satisfied. We can then avoid the awkward result of the above reasoning by adapting it in the following way:

1. It is a good thing now to act as if this teapot were filled with tea.[11]
2. By holding the teapot over the cup one acts as if it were filled with tea.

Therefore: hold the teapot over the cup.

The reasoning process here is still simple enough to make it plausible that a 2-year-old child could form an intention to pretend in this way. In a certain sense it is even simpler than a purely behaviouristic explanation that takes the child to act as if her action took place in an imaginary situation. On such an account the child would want to fill the cup not in reality per se, but in a

[11] The restriction to the present context is necessary because it is generally not a good idea to use an empty teapot as if it were filled, as children certainly know.

possible world that she imagines. The teleological account avoids this complication and does not require the child to imagine the pot to be filled with tea. The child can take the pot as it actually is and use it with a different goal, *as if* the purpose of her action were to fill the cup with tea.

If this explanation is on the right track, one can use it with minor modifications also to explain how children come to understand pretence actions in others. No new cognitive abilities need to be introduced. The same teleological reasoning that underlies their own pretend actions will suffice to explain how children make sense of what others do in pretend play.[12] One only has to connect an action that is considered to be worth doing with a particular agent, i.e. to think that it is a good thing *for this agent* to do it. With this modification made, the reasoning may proceed in the same way:

1. It is a good thing now *for this person* to act as if this teapot were filled with tea.
2. By holding the teapot over the cup she/he acts as if it were filled with tea.

Therefore: she/he holds the teapot over the cup.

Here too the comparison with a behaviouristic explanation shows that the teleological account is in one respect simpler. Children need not think that the other person fills the cup in some imaginary world. Rather, a child can take the other person to pursue the simpler goal of acting *as if* the conditions for acting successfully obtained (i.e. as if the pot were filled with tea).[13]

I do not want to suggest, however, that we should prefer a teleological explanation of pretend play just because it is simpler than a behaviouristic account. A behaviourist may respond here that his account is simpler in other respects. What else, then, would speak decisively in favour of a goal-centred explanation? I now want to argue that such an explanation has the extra power to explain cases that are beyond a behaviouristic theory's scope.

We have already seen that behaviouristic theories predict that children will distinguish pretending from trying only if the gestures and verbal expressions accompanying a certain movement have a straightforward meaning for them. It is clear for a child that scribbling with an expression of frustration is appropriate if one seriously *tries* writing with a pen, while making the same movements with an expression of satisfaction is appropriate if one *pretends* writing. But what about the case in which a person packs fruits into her lunch box and merely signals with a smile that she knows that these are not real fruits? How could children disambiguate her smile and thereby find out that she is merely pretending to prepare her lunch? In this case, the behaviouristic theory has nothing to offer for solving this problem. If we apply the teleological theory to this case, however, we can say that children learn to disambiguate a smile when they learn that it can be a sign for two different goals. The prediction would be that there is an age-difference to be observed in this respect: While children who already reason teleologically should be sensitive to the different meanings that a gesture can have, younger children would need less ambiguous behaviouristic clues like signs of success or frustration for distinguishing an act of pretence from an act of serious trying.

Further empirical work will be needed to show if children actually use their teleological reasoning capacities in this way. Apart from putting the theory to an empirical test, however, we may also consider its advantages at a theoretical level. As the contributions in this volume show,

[12] Strictly speaking, it cannot be the same reasoning process because in one case it has to produce the *command* to act in a certain way, while in the other case it has to produce a *description* of what the other person does. Some further apparatus will therefore be needed to explain this functional difference.

[13] This simplification need not imply, however, that children can pretend (and understand pretend acts) without employing a mechanism for 'quarantining' propositions. Again, I set aside here the question what functional architecture is needed for implementing this form of teleological reasoning behaviour.

philosophers and psychologists find it difficult to say exactly what makes a cognitive ability metacognitive and how we can recognize such abilities either in early childhood or in the behaviour of non-human animals. It is an intriguing idea that a teleological theory could make an important contribution to this debate since it steers a middle course between a behaviouristic and a mentalistic explanation of pretend play.

The experience of pretending: a metacognitive feeling?

In contrasting mentalistic and behaviouristic explanations of pretend play one soon reaches a point where the debate ends in an impasse. This is what the comparison between Leslie's theory and the behaviouristic theory of Nichols and Stich has revealed. From Leslie's point of view, children need a metarepresentational capacity before they can act as if a certain state of affairs obtained: they must be able to represent the belief that a certain state of affairs does not obtain and at the same time also represent the attitude of pretending that it obtained. Behaviouristic theories consider this to be an unreasonable demand, and suggest an explanation of pretence that works without attributing such metarepresentational capacities to children. Although Leslie rejects these behaviouristic theories as unsatisfactory, he concedes that children need not employ a *full-blown* representational theory of mind in order to meet the requirements of his theory. As we have seen, however, there is a stumbling block for both of these generic theories. Children distinguish trying from pretending before they reach the age at which they acquire a representational theory of mind, and they might be able to recognize this distinction even if the only sign of pretence given as a clue was a knowing smile. Neither theory can cope with this problem adequately, as I have argued in earlier sections.

The teleological theory carries the hope of showing us how to escape from this impasse. Though it is a different theory, it may retain some of the plausible ideas that can be found in the more radical theories that it purports to supersede. In the preceding section, I discussed both similarities and differences between a teleological and a behaviouristic explanation of pretence. I now suggest that a teleological explanation of pretend play could also accommodate certain elements of a mentalistic explanation. In particular, I will argue that it enables one to integrate Leslie's distinction between the full capacity of a representational theory of mind and a form of metarepresentation 'light' that he calls 'M-representation' in order to accommodate the idea that metacognitive abilities develop in stages. That idea is plausible independently of whether it can save Leslie's theory of pretend play.

How could this idea be integrated into a theory of pretence based on a teleological conception of intentional action? The claim to be justified is that *some* form of metacognition manifests itself in early childhood in *some* instances of pretend play. It need not be a full-blown metarepresentational ability, and it may manifest itself not in all occurrences of pretend play, but only when children's competence becomes more refined. Children that disambiguate a knowing smile serve as a paradigm of such cases. Could it be that children succeed in such tasks because they can rely on a metacognitive feeling that pretend play generates?

To get this project off the ground, we must first reject the presumption that metacognitive abilities are developmentally tied to mental concepts provided by a theory of mind. As long as this assumption is in place, any connection between pretence and metacognition would probably depend on accepting some version of Leslie's mentalistic theory of pretence. So we would still be caught in the problems created by such theories, and immediately face the behaviouristic objection that since mental concepts are not needed in early pretend play, metacognition could not play any role either. We can escape this impasse only by restricting ourselves to a form of metacognition that is *not* tied to conceptual abilities provided by a theory of mind.

Escaping this impasse requires us to take into account the role that metacognitive feelings may play in this context. In defining such feelings we need to invoke both functional and phenomenological terms. For instance, one could define metacognitive feelings as processes of monitoring and evaluating the performance of our own cognitive system, which make us experientially aware of information carried by certain internal states. Examples of such feelings are: the feeling of knowing something, the feeling of being able to remember something, the feeling of not knowing or not being able to remember something, feelings of certainty, feelings of uncertainty, etc. Psychologists investigate such feelings often in connection with metacognitive judgements that children can express verbally (see Dunlosky and Metcalfe 2009, chapter 4). This connection is made for good methodological reasons. Yet it must not mislead us into thinking that metacognitive feelings are somehow constitutively linked to judgements that articulate these feelings with the help of a particular conceptual framework. Feeling and judgement may be as different in this case as they are in cases of simple perceptions. It is one thing to have the experience of tasting a ripe strawberry, and a different thing to form the judgement, 'That is the taste of a ripe strawberry'. One needs the appropriate concepts for making this judgement, but one does not need them to enjoy the taste. For similar reasons, metacognitive feelings can be 'non-conceptual' in the following sense: they can draw attention to an internal state in terms of an experiential quality that this state exhibits prior to forming a judgement about what this quality consists in.[14]

Let us now return to the main question. Could certain forms of pretend play indicate the presence of metacognitive abilities because they are connected with a distinct type of *feeling* that arises from such abilities? To see the merits of this proposal, three questions need to be addressed. What could these metacognitive feelings be? What function could they have in children's pretend play? And how do we know that metacognitive abilities are involved in producing such feelings? Let us briefly consider each of these questions.

When the question is what feelings pretend play could produce, the natural place to start is with experiences with which one is familiar. Such a first-personal approach bears the risk of treating subjective experiences as data that others can check for correctness. It also creates the problem mentioned earlier in the first section, that we may project experiences onto other agents, and especially young children, that we may not actually share with them. Despite these significant methodological problems, however, the first-person approach remains the only possibility for describing experiences in the absence of contentious theses about the mechanisms that give rise to such experiences. This is the great advantage of phenomenological descriptions, which we must respect even if we have to be extremely careful about generalizing from first-person descriptions.

Phenomenological reflection leads me to believe that the most distinctive feeling one has while pretending is a sense of *freedom*. Pretend play enables one to engage in actions without concern for the consequences that these actions would normally have. For instance, the thrill one experiences in using a flight simulator partly depends on this feeling of freedom. It feels great to press the button for take-off and to land the plane safely because one knows that one is not risking one's life in doing so. As adults we tend to depend on such virtual simulators to obtain a feeling of freedom, whereas children seem to experience such freedom in much more mundane situations. For a child, it may feel just as great to lift an empty teapot above a cup as it does for the adult to successfully land a plane in a flight simulation. In this way, the child can engage in many actions that would otherwise be quite dangerous or inappropriate. She may experience freedom in pretending to fill a cup with tea without having to be afraid of spilling the tea or burning her hands.

[14] In a similar vein, one can distinguish 'declarative' and 'procedural' forms of metacognition. See Proust 2007 and chapter 14, this volume and Esken (Chapter 8, this volume) for pushing the same idea.

Having identified a feeling that is distinctively associated with pretend play, we can now ask what the function of having such feelings may be. An epiphenomenalist might claim that these feelings are just positive side effects of such playful behaviour, which serve no particular function at all. But such epiphenomenal explanations underestimate the motivating power that experiences often have. It is therefore likely that the thrill that children—and sometimes also adults—experience in pretend play is part of the pleasure they seek and makes it attractive for a child to engage in such play. This hypothesis fits perfectly with a teleological explanation according to which children consider it to be a *goal* to act as if things were different from what they are. It is not a trivial matter why children should find this particular goal attractive. Consider again the case of pretending to make a drawing with a pen that does not work. Children may do this simply because they enjoy 'scribbling' with the pen whether or not it leaves any marks on the paper. That is to say, they may enjoy the movement itself, but not the purpose with which this movement is normally performed. Hence it is difficult to say exactly what a child is doing when she 'scribbles' and expresses delight in doing so. Does she just enjoy the movement itself, or does she pretend that the pen is working fine? The answer depends on whether it is her *goal* to behave as if she were drawing a picture. Only if she sets herself this goal will her action generate more than the simple joy of scribbling, namely the more interesting feeling of 'writing' and 'drawing' without having to care about what the graphics produced look like. The upshot is that we need to take into account the feeling of freedom in order to explain why children should find the goal of pretending attractive. Otherwise the best explanation for their behaviour could be that they enjoy the movement that we then mistakenly interpret as an act of pretence.

This motivating explanation need not exhaust the function of pretend feelings. They may not only be crucial in *forming* an intention to pretend, they can also be important for *recognizing* an act of pretence performed by others. How could this be? If children were operating on the conception that other persons' intentions derive from their subjective beliefs and desires, their own experience in pretending would not matter much in comprehending such actions performed by others. However, if children regard pretending as a common goal that everyone finds attractive, these experiences clearly matter. Without the appropriate guidance by their own feelings, there would be no goal that children could grasp as a goal that others find attractive too. Thus, the experience of freedom could forge a link between the goal of pretending and the behaviour they observe in others when they pretend to do something.

That such a link exists is not pure speculation. The experience that children enjoy in pretending could explain certain mistakes that children make when they seem to follow a 'default-principle' in classifying behaviour as pretence. The fact that even much older children tend to use the concept of pretence quite indiscriminately points in this direction. Angeline Lillard showed this in a series of experiments in which she introduced 4- to 5-year-old children to a troll named Moe who hopped like a rabbit (Lillard 1993, 1994). Before the children observed Moe hopping like a rabbit, the experimenters told the children that Moe knows nothing about rabbits, including how they move. This information did not prevent most 4- to 5-year-old children from answering 'yes' when they were asked whether Moe is *pretending* to hop like a rabbit. They simply ignored the fact that one can pretend doing something only if one knows how to do it. Lillard takes this data to support a behaviouristic theory according to which children understand the question 'Does Moe pretend to hop like a rabbit?' as saying no more than 'Does Moe *behave* like a rabbit?' However, there is also another explanation why children make these mistakes. It also may be that children use the following default principle:

> If an agent acts as if he were doing *F*, although he is not in a position to do *F*, then he is pretending to do *F*.

This alternative explanation allows us to say that children do not simply ignore the information that Moe neither is a rabbit nor knows anything about rabbits. They may use this information to conclude that Moe is hopping like a rabbit despite not being in a position to do so. So they conclude with the above principle that he is pretending to act like a rabbit.

If this explanation is on the right track, then we have empirical evidence that children rely on their own experiences when attributing pretence to others. Otherwise it would be hard to explain why children should rely on such a simplistic principle if even much younger children already discriminate between acts of pretence and other forms of behaving-as-if. The key to such an explanation could be the joyful experience of pretending. Seeing Moe hopping like a rabbit, children may immediately associate with this observation the fun of playing a rabbit without having to be one. If they project this motivation into Moe's behaviour, they might misclassify it as a case of pretence for that reason.

This brings me to the final question, which is also the most difficult one. Why should the feelings that pretend play generates be *metacognitive* feelings? The difficulty here is that we know too little about metacognitive feelings as a general phenomenon. There are, of course, well-developed theories that deal with feelings of knowing and closely related phenomena (see Koriat 1993, Chapter 13, this volume). How far these theories also apply to other kinds of metacognitive feelings, such as feelings of freedom, is unclear. Some psychologists have pointed out that there is a strong connection between the development of self-awareness and the emergence of pretend play in early childhood (see Lewis and Ramsay 1999, 2004). Similar connections have also been investigated between imitation skills and pretend play (see Nielson and Dissanayake 2004). However, these correlations are not as informative as they seem, since the development of self-awareness is itself a controversial topic. While Ramsey and Lewis think that 2-year-old children already possess a well-developed self-concept that includes awareness of their own mental states, others would argue that children's self-concept at this age is still grounded mostly in bodily self-experience and not developed enough to include any metacognitive feelings (see Povinelli 1995).

In view of such diverging opinions, I think that the only promising route for establishing a link between pretend play and metacognition is the social route. The starting point here can be the teleological claim that children at a very early age already begin to grasp the intentional structure of pretend play. The second step is to point out that there are two sides to this glimpse into intentional structure: children know how to form an intention to pretend, and they recognize when others act with such an intention. As a third step, we can integrate into the theory the assumption that children experience a certain feeling of freedom when pretending. Now the question arises, could this experience of freedom make children *aware* of their intention to pretend, or do they simply enjoy acting accordingly? As long as we focus only on the behaviour of children in pretend scenarios, it will be difficult to find a reason for saying that these children are now *aware* of what they are intending to do, namely to pretend. The simpler assumption that they are aware of an 'objectively' attractive goal will be sufficient to explain their pretend behaviour. However, when we take into account their social competence, the following possibility gains some plausibility. For when children recognize that others pursue a goal that they also consider to be attractive, their social awareness may alter the way in which they experience their own behaviour. More precisely, the ability to recognize pretend acts in others may draw their attention to their own intentions when pretending. The result would be the metacognitive feeling of having such an intention. Children would thereby not become competent 'mindreaders', but rather would rely on their teleological conception of pretend play in interacting with other agents. Their growing self-awareness could make them more sensitive to what other agents have in mind, and thus enable them to recognize pretence even in cases when a knowing smile is the only behaviouristic

clue on which they can rely. The teleological approach may thus give rise to a *metacognitive feeling theory of pretence*.

Conclusion

I have argued in this paper that—for different reasons—mentalistic and behaviouristic explanations of pretend play are unsatisfactory. While behaviouristic theories underestimate children's early competence by severely limiting their ability to distinguish pretending from other forms of behaving-as-if, mentalistic theories tend to overestimate children's competence by granting them an insight into the psychological conditions of pretending. The teleological theory steers a middle course in sharing with the behaviouristic theory the claim that young children do not need any mental concepts for understanding simple acts of pretence, while on the other hand insisting that it takes more than a behaviouristic rule to grasp the goal of acting as if. Once we ask why children might find this goal attractive, we see that a positive experience in pretending is needed to rule out the problematic idea that children simply enjoy certain movements when engaged in acts of pretence. More evidence is needed, however, to decide when and how experiences of pretending may involve metacognitive feelings, including that of freedom. From a theoretical point of view, the best bet seems to be that the social competence that children manifest in recognizing an intentional act of pretence changes their own experiences of pretence. Whether such a feedback exists remains to be seen.[15]

Acknowledgement

This paper was written as part of the European Science Foundations EUROCORES Programme CNCC, and was supported by the Austrian Science Fund (FWF), project I94-G15.

References

Bloom, P. and German, T. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77, B25–31.

Csibra, G. and Gergely, G. (1998). The teleological origins of mentalistic action explanations: a developmental hypothesis. *Developmental Science* 1(2), 255–9.

Dunlosky, J. and Metcalfe, J. (2009). *Metacognition*. Los Angeles, CA: Sage Publications.

Friedman, O. and Leslie, A. M. (2007). The conceptual underpinnings of pretence: Pretending is not 'behaving-as-if'. *Cognition*, 105, 103–24.

Gergely, G. and Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Science*, 7(7), 287–92.

German, T. P. and Leslie, A. M. (2001). Children's inferences from 'knowing' to 'pretending' and 'believing'. *British Journal of Developmental Psychology*, 19, 59–83.

Harris, P. (1994). Understanding pretence. In C. Lewis and P. Mitchell (Eds.) *Children's Early Understanding of Mind: Origins and Development*, pp. 235–259. Hove: Lawrence Erlbaum Associates Ltd.

Harris, P. and Kavanaugh, R.D. (1993). Young children's understanding of pretence. *Monographs of the Society for Research in Child Development*, 58 (1).

Jarrold, C., Carruthers, P., Smith, P. K., and Boucher, J. (1994). Pretend play: Is it meta-representational? *Mind and Language*, 9(4), 445–68.

[15] I am grateful to several referees and readers who gave me feedback on earlier drafts of this chapter, including Frank Esken, Paul Harris, Josef Perner, Johannes Roessler and Daniel Siakel. Daniel also invested his time into improving my English for which he deserves special credit.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100(4), 609–39.

Leslie, A. M. (1987). Pretence and representation: The origins of 'theory of mind'. *Psychological Review*, 94, 412–26.

Leslie, A. M. (1988). Some implications of pretence for mechanisms underlying the child's theory of mind. In J. W. Astington, P. Harris, and D. R. Olson (Eds.) *Developing Theories of Mind*, pp. 19–46. Cambridge: Cambridge University Press.

Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition*, 50, 211–38.

Lewis, M. and Ramsay, D. (1999). Intentions, consciousness, and pretend play. In P. D. Zelazo, J. W. Astington, and D. R. Olson (Eds.) *Developing Theories of Intention. Social Understanding and Self-Control*, pp. 77–94. Mahwash, NJ: Lawrence Erlbaum Associates Ltd., Publishers.

Lewis, M. and Ramsay, D. (2004). Development of self-recognition, personal pronoun use, and pretend play during the 2nd year. *Child Development*, 75(6), 1821–31.

Lillard, A. (1993). Young children's conceptualization of pretence: Action or mental representation state? *Child Development*, 64, 372–86.

Lillard, A. (1994). Making sense of pretence. In C. Lewis and P. Mitchell (Eds.) *Children's Early Understanding of Mind: Origins and Development*, pp. 188–205. Hove: Lawrence Erlbaum Associates Ltd.

Lillard, A. (1998). Wanting to be it: Children's understanding of intentions underlying pretence. *Child Development*, 69, 981–93.

Lillard, A. (2001). Pretend play as twin earth: A social-cognitive analysis. *Developmental Review*, 21, 495–531.

Nichols, S. and Stich, S. (2000). A cognitive theory of pretence. *Cognition*, 73, 115–47.

Nichols, S. and Stich, S. (2003). *Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Clarendon Press.

Nielsen, M. and Dissanayake, C. (2000). An investigation of pretend play, mental state terms and false belief understanding: in search of a meta-representational link. *British Journal of Developmental Psychology*, 18, 609–24.

Nielsen, M. and Dissanayake, C. (2004). Pretend play, mirror self-recognition and imitation: a longitudinal investigation through the second year. *Infant Behaviour and Development*, 27, 342–65.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: The MIT Press.

Perner, J. (1995). The many faces of belief: Reflections on Fodor's and the child's theory of mind. *Cognition*, 57, 241–69.

Perner, J. (2004). Wann verstehen Kinder Handlungen als rational? In H. Schmidinger and C. Sedmak (Eds.) *Der Mensch—ein 'animal rationale'? Vernunft - Kognition—Intelligenz*, pp. 198–215. Darmstadt: Wissenschaftliche Buchgesellschaft.

Perner, J. and Roessler, J. (2010). Teleology and causal understanding in children's theory of mind. In J. Aguilar and A.A. Buckareff (Eds.) *Causing Human Action: New Perspectives on the Causal Theory of Action*, pp. 199–228. Cambridge, MA: The MIT Press.

Perner, J., Baker, S., and Hutton, D. (1994). Prelief: The conceptual origins of belief and pretence. In C. Lewis and P. Mitchell (Eds.) *Children's Early Understanding of Mind*, pp. 261–86. Hove: Lawrence Erlbaum.

Piaget, J. (1945/1962). *Play, Dreams, and Imitation in Childhood*. New York: Basic Books.

Povinelli, D. (1995). The Unduplicated Self. In P. Rochat (Ed.) *The Self in Infancy. Theory and Research*, pp. 161–92. Dordrecht: Elsevier Science B.V.

Proust, J. (2007). Metacognition and meta-representation: is a self-directed theory of mind a precondition for metacognition? *Synthese*, 2, 271–95.

Rakoczy, H., Tomasello, M., and Striano, T. (2004). Young children know that trying is not pretending: A test of the 'behaving-as-if' construal of children's early concept of pretence. *Developmental Psychology*, 40(3), 388–99.

Rakoczy, H. and Tomasello, M. (2006). Two-year olds grasp the intentional structure of pretence acts. *Developmental Science*, 9(6), 557–64.

Ristau, C. A. (1991). Aspects of the cognitive ethology of an injury-feigning bird, the piping plover. In C. A. Ristau (Ed.) *Cognitive Ethology. The Minds of Other Animals*, pp. 91–126. Lawrence Erlbaum.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13(1), 103–28.

Chapter 10

The development of earlier and later forms of metacognitive abilities: reflections on agency and ignorance

Daniela Kloo and Michael Rohwer

Metacognition encompasses both 'metacognitive knowledge', that is, knowledge about our own mind, and 'metacognitive regulation', that is, the regulation of one's cognitive activities (Moses and Baird 1999). 'Metacognitive regulation' can be seen as part of (but not equated with, see Perner Chapter 6, this volume) executive functioning, which refers to the conscious regulation of one's cognitions, actions, and emotions, with metacognitive regulation referring to the conscious regulation of one's cognitions. And, 'metacognitive knowledge' can be seen as part of theory of mind, which refers to the ascription of mental states (e.g. intentions, beliefs, and desires) to oneself (metacognitive knowledge) and to other people.

The development of these metacognitive abilities spans several years. For example, some aspects of *metacognitive regulation* show developmental improvements from childhood to adolescence. The ability to flexibly shift mental set considerably improves during preschool (e.g. Zelazo et al. 1996; Kloo et al. 2010), but switching abilities continue to increase from childhood to young adulthood (e.g. Cepeda et al. 2001). Also, planning abilities continue to develop from preschool well into early adolescence (e.g. Luciana and Nelson 1998).

In this chapter, however, we will focus on the development of *metacognitive knowledge*. First, we will concentrate on a rather early form of metacognitive knowledge, children's 'sense of agency'. Then, we will elaborate on later developments in children's 'knowing about knowing'. Finally, we will conclude that empirical evidence in both these domains of metacognition suggests that we have to thoroughly distinguish—theoretically and empirical research—between pre-reflective and reflective forms of metacognition (cf. Gallagher and Zahavi 2008).

The 'sense of agency'

The 'sense of agency' can be defined as 'the sense that I am the one who is causing or generating an action. For example, the sense that I am the one who is causing something to move, or that I am the one who is generating a certain thought in my stream of consciousness' (Gallagher 2000, p. 15). Gallagher (2010) distinguished this pre-reflective, first-order (minimal) sense of agency (SA1) from a reflective higher-order sense of agency (SA2), which involves the explanation of one's actions in terms of underlying intentional states.

Our (pre-reflective) 'sense of agency' is a rather well-ingrained capacity. For example, Metcalfe and Green (2007) showed that, generally, people are well able to judge whether they are controlling the movement of an external object (e.g. the position of a cursor in a computerized task) or not, when their control is limited by random noise. Metcalfe et al. (2010) showed that school-aged (8- to 11-year-old) children are also sensitive to their lack of agency in this task.

Unfortunately, we know little about the development of this pre-reflective sense of agency (SA1) in preschool children. Although SA1 should develop before SA2, most studies investigating the sense of agency in preschool children focused on their understanding of the mental states underlying agency, such as intentions. That is, most studies investigated children's reflective sense of agency (SA2).

In general, intended actions can be distinguished from accidental actions like mistakes or reflexes by comparing one's intention and the actual outcome. A number of studies indicate that 4-year-old children do not yet fully understand the nature of intentional actions, because they have difficulties understanding the distinction between desires and intentions in situations where goals and outcomes do not match.

For example, Schult (2002) showed that 3- to 4-year-old children, in contrast to 5- and 7-year-old children, have problems differentiating desires and intentions when presented with scenarios, in which intentions and/or desires were satisfied or were not. Similary, Phillips et al. (1998) found that 4-year-olds (compared to 5-year-olds) as well as young people with autism have difficulty distinguishing intended from unintended outcomes.

In their study, participants played an electronic target-shooting game and were thereby given first-hand experience of intended vs. accidental action outcomes. Critically, in this task, intentions and desires were separated. Participants had to shoot down coloured cans balanced on a wall, some of which contained prizes. That is on each trial, participants had (1) the desire to win a prize and (2) the intention to hit a particular can. However, which target can actually fell from the wall and whether it contained a prize or not was controlled by the experimenter. So, there were four possible outcomes. In two conditions, both the intention and the desire were either fulfilled or frustrated. In two conditions, the conditions of interest, there was a discrepancy between desire and intention. In one condition, the intention was fulfilled, but the desire was frustrated (the child hit the intended can, but it did not contain a prize). In the other condition, the desire, but not the intention, was fulfilled (the child did not hit the intended can, but this can contained a prize). After each trial, children were asked, 'Which colour did you mean to shoot?'.

In general, even the children with autism and the 4-year-olds showed some understanding, when asked to judge their intention. However, 4-year-olds (compared to 5-year-olds) had some difficulty in both miss conditions and in the discrepant hit condition. When they hit the intended can, but it did not contain a prize, they claimed to have intended to hit another can; however, as discussed by Philips et al., this may have been due to pragmatic difficulties.

Also, participants with autism were less accurate than a mental handicap control group in both miss conditions (when they did not hit the intended can) and over-attributed intentionality, but they had no problems in the hit conditions. This fits with similar findings on action-monitoring problems in children with autism (Russell and Jarrold 1998).

Other studies investigating children's understanding of reflex movements suggest that 3-year-olds also are unable to understand the non-intentionality of accidental reflex movements: for example, they have problems to correctly judge whether they 'meant to' produce a reflex movement as shown in a study by Shultz and colleagues (1980). These authors induced a knee-jerk reflex in 3- to 5-year-old children. Interestingly, when asked whether they meant to move their leg, 3-year-olds claimed that they meant to move their leg, but 5-year-olds correctly stated that they did not mean to move their leg.

According to Perner (1991), in order to understand that involuntary reflex movements are not intentional, children need to understand mental states as representations with causal efficacy. Indeed Lang and Perner (2002) found a relation between understanding reflex movements and understanding that (false) mental representations are causally responsible for our actions. They elicited the knee-jerk reflex in 3- to 5-year-old children and asked them: 'Look your leg moved!

Did you mean to do this?' About 50% of the children incorrectly claimed that they meant to move their leg. Interestingly, understanding the involuntary nature of reflexes in the knee-jerk reflex task was strongly correlated (even when age and verbal intelligence were partialled out: $r = 0.60$) with tasks involving situations where false mental representations lead to mistakes (false belief task and dimensional change card sorting task).

This fits with a host of studies showing that there is a robust correlation between theory of mind and executive functions in typically developing children aged 3–5 years (for a review, see Perner and Lang 1999). More specifically, a relation between various theory of mind measures (e.g. the false belief task) and conflict tasks like the Bear/Dragon task, the Whisper task (Kochanska et al. 1996), the Dimensional Change Card Sorting task (Zelazo 2006), or Luria's hand game (Hughes 1998) has been repeatedly demonstrated even with age and verbal intelligence partialled out (e.g. Frye et al. 1995; Carlson et al. 2002; 2004; Lang and Perner 2002; Perner et al. 2002). Conflict tasks typically require the suppression of an inappropriate, often prepotent, response and the simultaneous activation of a conflicting response.

In sum, young preschool children do not yet fully understand the nature of mental states, such as intentions, that are related to our sense of agency. They tend to conflate intentions and desires (e.g. Schult 2002) and they do not understand the non-intentionality of accidental reflex movements (e.g. Shultz et al. 1980). That is, 3- to 4-year-old children seem to lack a reflective higher-order sense of agency (SA2). In turn, their immature reflective sense of agency may be related to difficulties on theory of mind measures and executive function tasks.

But what about their pre-reflective, first-order (minimal) sense of agency (SA1)? Do 3-year-olds have a first-order sense of agency? And how is SA1 related to theory of mind and executive function abilities?

In a recent study (Kloo and Pinnitsch 2010) with 101 3- to 4-year-old children, we investigated this question. We focused on one example of Gallagher's (2000) definition of SA1: the sense that I am the one who is causing something to move. To this end, we invented a racecourse-game for measuring the SA1 in 3- to 4-year-old children by focusing on their action-monitoring abilities. In our electronic racecourse-game, children had to decide which one of two cars (a white one and a red one) they are controlling with their joystick. One of the cars was controlled by the experimenter; the other car was controlled by the child. After the race had started, children had to state, as quickly as possible, which colour their car was.

We found that even young 3-year-olds (37–41 months old) were quite good at this task; 64% correctly identified 'their' car on two or three of the three test trials. And they were almost as good as 4-year-olds (50–57 months old), 75% of whom passed two or three test trials. In addition, performance on the racecourse-game did not improve in this age range. This suggests that even 3- and 4-year-old children are considerably (though not perfectly) certain about their own actions and do experience themselves as an agent.

In addition, children's 'sense of agency' (as measured by the racecourse-game) was significantly related ($r = 0.29$) to their understanding of their own mental states (as measured by their ability to recall their own earlier false belief). However, performance on the racecourse-game was not significantly related ($r = -0.07$) to children's understanding of other people's mental states. Also, children's ability to monitor their actions in the racecourse-game was not significantly related to performance on executive function tasks (Dimensional Change Card Sorting task: $r = -0.05$; Luria's tapping task: $r = -0.09$). However, performance on the executive function measures was significantly related to understanding other people's mental states ($r = 0.29$).

Given the fact that even young 3-year-olds performed quite well on our 'sense of agency' task and that children's 'sense of agency' was related to their ability to recall their own earlier false belief, this suggests that a pre-reflective 'sense of agency' may be a precursor for understanding one's own

mental states. Interestingly, understanding *other* people's mental states was related to executive control abilities—but not to understanding one's own mental states. This suggests that perspective-taking is more strongly related to 'metacognitive regulation' (the regulation of one's cognitive activities) than to metacognitive knowledge. For example, perspective-taking may be the result of adequate 'metacognitive regulation' or executive control (e.g. by inhibiting one's own perspective).

A similar dissociation between the 'sense of agency' and understanding other people's mental states has been found in persons with autism. David et al. (2008) found that high-functioning adults with autism showed significant mentalizing deficits but no deficit on an action-monitoring and action-attribution (to self vs. other) task. We, therefore, suggest that the 'sense of agency' (i.e. an awareness that I am the initiator of an action) is a precursor for later metacognitive knowledge, for example, for remembering one's own earlier false belief. In contrast, understanding other people's minds depends more strongly on metacognitive regulation (for example, on inhibiting one's own perspective), a domain in which persons with autism are known to have difficulty (see, for example, Hill, 2004, for a review).

Knowing about knowing

Another form of metacognition is children's understanding of their own epistemic states, that is, their understanding of their own knowing and not knowing. Given that such metacognitive (better meta-epistemic) insights are of fundamental importance for competent performance in a number of different domains of cognitive development (e.g. for successful communication, for reading and listening comprehension, for memory performance, etc.) it is not surprising that an increasing number of different methods have now been developed to study their ontogenesis.

Reflective metacognition

Knowing about own knowing and ignorance—verbal (and direct) approach

The most direct and most often used method to study metacognition in children is based on the tacit assumption that humans are aware of their epistemic states (and of the informational sources of these epistemic states) and can thus explicitly declare their meta-epistemic insights to others. In laboratory tasks, typically, children are asked metacognitive questions about their own epistemic states about a hiding event. A number of studies consistently indicate that children around three years acknowledge their knowledge of the contents of a container, when they had looked inside it (e.g. Pillow 1989; Ruffman and Olson 1989; Pratt and Bryant 1990; Tardif et al. 2005). However, a conflicting pattern of results exists as to when children can acknowledge their own ignorance, when they weren't allowed to look inside the container. For instance, in a study by Wimmer et al. (1988, experiment 1), 3- to 5-year-old children were shown various boxes. Two children were always sitting opposite to each other on a table on which various boxes with an unknown content were placed. The experimenter then either showed (visual access) or told (without visual access) one of the two children the content of the box, whereas the other child remained *totally ignorant about what the content could possibly be*. Children then had to assess their own epistemic state and the epistemic state of the other child. While only 50% of the 3-year-olds were able to correctly assess their own epistemic state, 94% of the 4-year-olds, and 100% of the 5-year-olds made correct assessments. Children from 4 years on therefore seemed to have few problems acknowledging their own ignorance (see also Pratt and Bryant (1990), for results that even 3-year-olds can acknowledge their own ignorance in a similar *total ignorance task*).

However, children experience more problems in assessing their own ignorance in a *partial exposure task*, in which they are exposed to a range of objects, but then cannot see which object is

being put inside a box. This was found to be the case in a control condition by Sodian and Wimmer (1987, experiment 1). In this experiment, 4-year-olds and 6-year-olds were shown a container with two kinds of differently coloured balls in it. Then one of the balls was covertly transferred to a bag by the experimenter. Children were only told that one of the two balls had been transferred to the bag and were then asked: 'Do you know what colour the ball in the bag is?'. About 35% of the 4-year-olds and even 13% of the 6-year-olds wrongly claimed 'to know'. In their experiment 2, even 50% of the 4-year-olds overestimated their knowledge of which ball had been transferred.

We (Rohwer et al. 2012a) explained these results—spanning an age range of several years for the onset of children's capacity to metacognitively acknowledge their own ignorance—with structural differences between the studies. That is, we argued that partial exposure tasks (e.g. Sodian and Wimmer 1987) are more difficult for young children than total ignorance tasks (e.g. Pratt and Bryant 1990), because children until 5 or 6 years do not yet understand that their knowledge has causal origins (and thus stems from an informative access given at the right time). Instead, preschool children rely on a *sense of knowing* when assessing their epistemic states. That is, when young children are asked a metacognitive knowledge question like 'Do you know what is in the box?' they just check whether they can *easily* think of some plausible object name and if so they have a sense of knowing and answer affirmatively ('Yes, I know') to the knowledge question. Otherwise they deny having knowledge ('No, I do not know'). Thinking of a plausible object is easier when one has been exposed to a set of potential objects to be hidden (partial exposure task) than in a task in which nothing relevant is shown (total ignorance task). Hence the young children are more likely to answer wrongly with 'I know' in the partial exposure than in the total ignorance task.

We tested this explanation with 3- to 7-year-olds comparing their epistemic state assessments on a total ignorance task, a partial exposure task (two toys were shown ahead of hiding), and a complete knowledge task (children were allowed to watch how a single object was 'hidden' in a container). After the toy had been hidden in the container, the experimenter said (in all of the tasks): 'Do you know now which toy is inside or do you not know?' In line with our explanation, evidence was found that all age groups had no problems in assessing their own knowledge correctly after having been allowed to watch the hiding. The vast majority of the 3-years-olds (97%) could additionally report their state of ignorance accurately in the total ignorance task. However only about 30% of the children before 6 years could correctly acknowledge their own ignorance in the partial exposure task. The overwhelming majority of children before 6 years thus wrongly claimed to 'know' on the test question and additionally re-affirmed their knowledge statement on a subsequent 'know-guess' control question ('Do you really know that or are you just guessing?'). In experiment 2 of this study further evidence was obtained that children under the age of 5 also over-estimated their knowledge in partial exposure tasks, independently of whether 2, 3, 5, or 10 potential objects were shown to them ahead of hiding. Taken together, we thus found evidence for our assumption that children before 5 or 6 years of age do have a wrong conceptualization of knowledge, that is, they do not yet understand that knowledge has causal origins and instead rely on the *ease* with which plausible information comes to their mind (i.e. they rely on a *sense of knowing*) when assessing their epistemic states.[1,2]

[1] Noteworthy, the ease with which information comes to mind (e.g. Kelley and Lindsay 1993; Koriat 1993; Mazzoni and Nelson 1995) and the ease with which information is accessible or the efforts experienced in reaching a decision (e.g. Kelley and Lindsay 1993; Nelson and Narens 1990; Zakay and Tuvia 1998) have also been found to influence (and strengthen) adults' subjective confidence in (e.g.) the correctness of retrieved information in memory tasks.

[2] Note that there is also an interesting link to other developmental evidence (Beck et al. Chapter 11, this volume) as the ability to easily imagine an outcome of a task also seems to play an important role in the

Supporting evidence that children before the age of 6 years do indeed have a somewhat deviant conceptualization of knowledge also comes from so called *know-guess tasks*. In these tasks, the experimenter hides an object in one of two containers. The child is then asked to indicate the container with the object. Not knowing (or having seen) where the object is the child can only guess. In cases of successful retrieval the younger children tend to answer the question 'Did you really know where the object was or did you just guess?' wrongly with 'known.' This confirms that young children identify knowing with getting it right without concern for the causal origins of knowledge (i.e. without concern for knowledge's necessary evidential basis; Miscione et al. 1978; Johnson and Wellman 1980; Perner 1991; Perner and Ruffman 1995).

Understanding the causal origins of own knowledge—verbal (and direct) approach

Other findings in the literature also support the view that children before 5 do not understand the causal origins of their own knowledge. Gopnik and Graf (1988) for instance asked children to report how they had learned about the contents of a container (i.e. by looking inside, by being told about the contents, or by figuring them out on their own by means of a clue). Three-year-olds often could not identify the source of their knowledge, claiming for instance to have seen an object in a drawer when they had in fact been told about it by the experimenter. By 5 years of age children were, however, able to justify their knowledge state by adequately reporting how they had learned the information in question.

Similarly, in a study by Taylor et al. (1994, experiment 1), preschool children showed a striking neglect of the sources of their own knowledge as they could not identify when they had learned specific information. The majority of 4-year-olds (86%) and 5-year-olds (57%) who had been taught novel facts (e.g. that tigers' stripes provide camouflage) about animals, insisted to have known these facts for a long time, although the learning event in fact had taken place only a few minutes ago. Interestingly, such evidence was also obtained by Gopnik and Astington (1988) who found that young preschoolers claimed to have always known that a Smarties box contained pencils, although this piece of knowledge had in fact only been learned a moment before (after children had been allowed to look into the Smarties box).

Four- and 5-year-olds have also been reported to have an incomplete understanding of the specific type of knowledge which is to be gained from a specific sensory source, that is, children of this age cannot make an explicit judgement about the sense through which information has been acquired (e.g. young preschool children do not understand that feeling an object does not provide information about visual details of this object etc.; O' Neill and Gopnik 1991; O'Neill et al. 1992; Pillow 1993; O' Neill and Chong 2001).

Knowing about own knowing and ignorance—behavioural (and direct) approach

While it could be argued that linguistic demands could have contributed to the difficulty in some of these tasks, that is, children fail when verbal and explicit metacognitive judgements are required, children do not fare better when being allowed to express their own ignorance or uncertainty by means of a non-verbal rating scale or by means of an explicit behavioural choice. For instance 5-year-olds in a partial exposure task by Pillow (2002) who had only seen a pair of two differently

fact, that children overestimate their own knowledge more frequently in specified epistemic uncertainty tasks (in which an outcome can easily be imagined) than in unspecified epistemic uncertainty tasks or physical uncertainty tasks (in which it is more difficult to imagine an outcome).

coloured balls ahead of hiding, but then had not seen which of them was being hidden, rated their own statement about the colour of the hidden ball as overly certain on a non-verbal rating scale, thus failing to acknowledge their own uncertainty (77% out of 100% certainty in *guess task* of experiment 1; 82% out of 100% certainty in *informed guess task* of experiment 2).

Further support for the notion that children over-estimate their own competence even in behavioural partial exposure tasks comes from recent studies by Rohwer et al. (2010, 2012b) in which non-verbal response options were used. In one of these studies (Rohwer et al. 2010), children from 3–8 years of age, saw, for instance, two kinds of animals, a cat and a dog, one of which was then covertly transferred to an opaque building, the animal house. Children then had to decide whether they wanted to feed the unknown animal in the house with either a bone (which could only be eaten by the dog) or a fish (which could only be eaten by the cat) or whether they wanted to place both kinds of foods to the house in order to ensure that the animal, be it a cat or a dog, would have something suitable to eat. Children who placed both kinds of food to the house and thus acknowledged their ignorance or uncertainty behaviourally won a star, whereas children who over-estimated their knowledge lost one. It was found that until 6 years of age the majority of children still overestimated their own knowledge even when they were made subject to losses. Similar results were also obtained in a more narrative task context (Rohwer et al. 2012b), in which children from 3–8 years had to decide into which one out of three houses an unknown toy animal in an opaque box (which an animal catcher had covertly caught before) should be placed. Three houses were available: a cathouse containing only cats and food, a dog house containing only dogs and food, and an animal house containing only food. The child knew that the unknown animal in the box was either a cat or a dog and was told that the worst thing which could be done in this task was to place the animal in the box (e.g. cat) into the house inhabited by the opposite kind of animal (e.g. to put a cat into the doghouse), as then a fierce fight would start. The main question was whether children would play it safe and would place the opaque box with the unknown animal into the animal house to avoid making a possible error (thus acknowledging their own ignorance) or whether children would immediately place it to one of the other two houses, thus over-estimating their own competence. It was found that only by 6 to 7 years of age did children start to use the animal house as an option to avoid making a possible error. Younger children mainly opted for one of the other two houses.

That preschool children's difficulties are indeed not rooted in linguistic deficits (but in their reliance on a sense of knowing) is further supported by the fact, that preschool children were also found incapable of behaviourally expressing their own ignorance in many ambiguous referencing tasks. That is to say, when preschool children for instance do not see in which box an object has been hidden and are then given a message by the experimenter which equally refers to two or more possible hiding locations of the object (e.g. message refers to a large box when there are in fact two large boxes and one small box) then children do not act behaviourally in a way which would demonstrate that they acknowledge both possibilities and their ignorance (i.e. preschool children for instance place a marker to only one of the two large boxes but not to both of them; Robinson and Whittaker 1985; Robinson et al. 2006, experiment 2).

Knowing about own knowing and ignorance—behavioural (and indirect) approach

In other ambiguous referencing studies, a more indirect approach is used, which is based on the rationale that people can react adaptively to uncertainty or ignorance by either seeking additional information or deferring/delaying a response. However even with this different methodology children until 6 or 7 were often found to be oblivious to their own uncertainty. They failed to disambiguate ambiguous situations either by delaying a response briefly for obtaining more

information (Beck et al. 2008, experiments 1 and 2), or by lifting up cups (Beck and Robinson 2001, experiment 3), or by asking an experimenter (Ironsmith and Whitehurst 1978) or an 'informed man' in a game (who had seen into which house a puppet had gone) clarifying questions (Sommerville et al. 1979).

Interim summary

Many studies in the literature support the view, that children only acquire complete meta-epistemic competency relatively *late* in their childhood. That is to say, although children are able to correctly assess their own knowledge states, they have fundamental problems in being able to accurately gauge their own uncertainty or ignorance. These problems seem to be rooted in a wrong conceptualization of knowledge, that is, preschool children take a 'sense of knowing' for knowledge and strikingly neglect the causal origins of their own knowledge.

Pre-reflective access to own knowledge and ignorance

Behavioural (and indirect) approach

Importantly this is, however, not the complete picture which emerges from the empirical findings in the literature, as there are also studies which have yielded evidence that children do have an early sensitivity to their own ignorance or uncertainty. In a study by Call and Carpenter (2001, experiment 3) 27- to 32-month-olds showed some sensitivity to their own ignorance in a partial exposure task by seeking clarifying information before committing themselves to indicating where the object was. Stickers were hidden in one of *three* tubes, which the children were allowed to see before being hidden (partial exposure task). Children were then either allowed to watch the hiding (informative trial) or they were prevented from watching (partial exposure trial). After the hiding children could immediately select one of the tubes by touching or exploring the tubes before selection. In approximately 75% of the trials children explored the tubes by looking through them when they had not seen the hiding compared to about only 40% of the trials when they had seen the hiding. The authors interpreted this as indication that even very young preschool children have an early sensitivity to own ignorance or uncertainty.

Similarly, a recent study by Robinson et al. (2008) found that children's explorative behaviour reveals an earlier understanding of the modality specificity of knowledge than revealed in standard tasks, in which children have to make explicit verbal judgements about the sense through which information has been acquired (O' Neill and Gopnik 1991; O'Neill et al. 1992; Pillow 1993; O' Neill and Chong 2001). In Robinson et al.'s (2008) first experiment children had the task of identifying which one of a pair of toys the experimenter had placed on the table in front of them. For some pairs, both toys felt the same but differed in colour, for example a red and a blue cat. For other pairs, both toys looked identical, but they felt different, for example a hard and a soft bear. At the beginning of each trial, the child saw and felt both toys in a pair and agreed on their properties, for example, that they looked the same but felt different. Then the experimenter covertly mixed up the toys, and placed one on the table while asking 'Which one is it?' Interestingly, 3- and 4-year-old children were found to be more likely to touch the toy before answering when it was defined by hardness or softness, and to answer without touching it when it was defined by colour. That is, children seemed to understand when feeling was necessary and when seeing was sufficient. A second group of children played a similar game, but instead of placing the toy on the table, the experimenter gave it to the child so that the child saw and felt it at the same time. Children in this group were asked 'Which one is it?' and 'How did you know it was the (hard) one?' Although children always identified the toy correctly they were often unable to refer to the

correct sensory modality through which they had acquired their knowledge. Children's pre-reflective-/implicit information seeking behaviour thus revealed greater competence than their reflective-/explicit verbal judgements. This conclusion is underlined by a recent study by Balcomb and Gerken (2008). Even 3½-year-old children skipped uncertain memory trials in a recognition memory test in order to optimize their memory performance. The authors also interpreted children's performance as an indicator of an early implicit access to own ignorance or uncertainty.

That a pre-reflective-/implicit access to own ignorance or uncertainty might indeed precede children's ability to reflectively-/explicitly judge their own ignorance is further supported by evidence that even very young children, who normally fail in ambiguity tasks in terms of their explicit metacognitive judgements, can often exhibit implicit signs of uncertainty, like making more eye contact with a speaker, showing puzzled expressions, or showing prolonged response latencies when being confronted with an ambiguous input (e.g. Bearison and Levey 1977; Patterson et al. 1980; Flavell et al. 1981; Plummert 1996; Sekerina et al. 2004).

Pre-reflective access to own knowing and ignorance: lingering doubts?

The evidence for this pre-reflective- (or implicit) access view was however, often hard to interpret, as children in some studies failed to show evidence for such an early sensitivity in very similar task settings. For instance, while 2½-year-old children in the study by Call and Carpenter (2001) sought clarifying information in a partial exposure task, children before 6 or 7 years have consistently been found to fail to seek clarifying information in ambiguous referencing tasks (e.g. Beck and Robinson 2001, experiment 3). As we found this discrepancy in children's performance puzzling, we re-investigated whether children's behaviour would indeed differ between these two task types. In our study (Rohwer et al. 2012b), we thus compared a set of partial exposure tasks with a set of ambiguous referencing tasks. In our study children from 3–8 years were told that a star would be hidden under one of a couple of cups. Children were further told that their task was to find and lift the cup which contained the star. Children could either peek into all the cups before lifting the cup and committing themselves to a response or they could lift the correct cup straightaway. If an ambiguous trial was played the experimenter additionally gave the children an ambiguous message after the hiding of the star had taken place, like: 'I'll tell you now under which cup the star is, it is under the red one' (either referring to two red cups or to three red cups depending on the task). Show me under which cup the star is under.' If a partial exposure trial was played children only saw a set of two or three cups ahead of hiding and then the star was covertly hidden, followed by the same procedure as already indicated. There was also a set of control tasks (informative tasks) in which we checked whether children would lift the correct cup straightaway after having seen where we had hidden the star or after having been conclusively told about where it had been hidden.

Interestingly even 3- to 4-year-olds were found to seek clarifying information significantly more often in the inconclusive hiding tasks (both ambiguous referencing tasks and partial exposure tasks) than in those tasks in which they were either allowed to watch the hiding or in which they received conclusive verbal information about where the star was hidden. That is, children did not peek into the cups every time before they lifted them (although peeking into the cups before lifting would have been the most successful strategy for always winning stars), but frequently used the most appropriate (or adult-like strategy) in dependence on the visual- or verbal information received. In keeping with the fact, that preschool children normally fail in ambiguous referencing- or partial exposure tasks in terms of their explicit metacognitive judgements did we thus conclude that our findings can be taken to reflect a pre-reflective- or implicit access preschool children exhibited to their own ignorance.

Theoretical aspects: reflective metacognition versus pre-reflective access to own cognition

In order to explain why preschool children in some studies (e.g. Bearison and Levey 1977; Call and Carpenter 2001) were found to demonstrate an early implicit sensitivity to their ignorance, but failed to exhibit a comparable sensitivity to their ignorance in similar studies (e.g. Beck and Robinson 2001, experiment 3) we (Rohwer et al. 2012b) attempt a speculative account. We maintain that the discrepancy between the studies is caused by the fact, that children in some tasks react on the basis of a mere state of '*being ignorant*', whereas children in other studies attempt to '*represent their state of ignorance*' ('I know that I do not know') by means of explicit metacognitive reflection. We deem this discrepancy to be important because we assume that when preschool children attempt to metacognitively reflect on their epistemic state do they rely on what we call a wrong predictor of knowledge (i.e. a 'sense of knowing') and thus get trapped in over-estimating their competency.[3]

Summary

Taken together, when the literature regarding children's understanding of their own epistemic thinking is reviewed, there is a lot of evidence that children can only reflectively/explicitly assess their own epistemic states when they reach school age. The fact that preschool children's persistent problems are however even then limited to specific task settings, like partial exposure tasks, or ambiguous referencing tasks suggests a very specific deficit of metacognition of knowledge, in particular, an inability to metacognitively judge their own state of ignorance. Before they reach school age children are thus metacognitively ignorant about their state of ignorance, i.e. they are meta-ignorant. Only by about 5 to 6 years of age do they amend this deficit and, thus, are able to escape their meta-ignorance. Preschool children's meta-ignorance should, however, not be understood as a complete failure to monitor their own abilities. Their earlier competence in the total ignorance and complete knowledge tasks shows they must be able to reflect on their ability to produce something like a 'relevant guess', which gives them the feeling of being able to respond competently. However, they mistake this feeling as knowledge, as we have argued (Rohwer et al., 2012a) to explain their wrong affirmative answers in the partial exposure tasks. Older children and adults, too, often rely on their feeling of being able to produce a correct answer for judging their own knowledge. For instance, in the tip-of-the-tongue phenomenon we can be certain that we know the answer to a question, e.g. 'What is the capital of New York?' even if we cannot actually retrieve the answer. This feeling or 'sense' of knowing is deemed to be due to how easy our retrieval attempts feel (e.g. Koriat, 1993).[4] However, there is an important difference to what the younger children seem to be doing in many epistemic task settings, in which they are normally pre-destined to fail, like (e.g.) partial exposure tasks. Even if the first answer that rushes to our mind is 'New York City,' we do not announce it and claim that we know that it is NYC and firmly deny having guessed. For, it is not any plausible answer that comes to mind that we admit as knowledge. We must be able to find a suitable propositional fact like, 'Albany is the capital of

[3] An epistemic state sensitive responding (e.g. puzzled expressions, longer reaction times, or a correct information seeking behaviour) in young children should then consequently be seen as being directly triggered by the intensity of the epistemic state being present on the object level (to which even infants could have an implicit- or pre-reflective access.)

[4] See also Koriat (Chapter 13) in this book for a review of studies which have obtained evidence that retrieval fluency also has a significant impact on adults' judgements of learning (JOLs), feelings of knowing (FOKs), and their retrospective confidence judgements.

New York,' in our knowledge base before we can admit to having found the correct answer. In the case of the partial exposure tasks, there is no such relevant fact, e.g. 'the toy car is in the box,' in the child's knowledge base. The children do not check whether they have a definitive fact available for answering the question, they seem content with just any plausible answer they can easily muster. Their spontaneous guess seems to turn subjectively into knowledge. Mistaking their relevant guesses for actual knowledge constitutes a deep limitation in being able to accurately gauge their own ignorance.

Importantly however there is a flip side to this, as an early competence (or ignorance sensitivity) can be found nonetheless in preschool children as long as children in a task can act from a mere state of being ignorant, that is, as long as they do not have to make explicit/reflective judgements of their own epistemic states. As we have argued, early competence can be revealed in these tasks, as children do not attempt to rely on a wrong predictor of knowledge here (i.e. a 'sense of knowing') and thus avoid getting trapped in a state of meta-ignorance.

To conclude, research on the development on children's 'sense of agency' as well as research on the development of children's 'knowing about knowing' indicates that we clearly have to distinguish between reflective metacognitive abilities and pre-reflective cognitive skills. Without such a thorough distinction we might get trapped in a bulk of inconsistent findings, but by distinguishing between reflective and pre-reflective forms of (meta-)cognitive abilities we might be able to resolve many discrepancies found in the literature.

Acknowledgements

This study is part of two research projects financed by the Austrian and European Science Fund (FWF project V00-122 and ESF/FWF project I93-G15 'Metacognition of Perspective Differences'). The authors thank Andrea Pinnitsch, Laura Blank, and Monika Hildenbrand for help with data collection and coding.

References

Balcomb, F. K. and Gerken, L. A. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Science*, 11, 750–60.

Bayne, T. (2008). The phenomenology of agency. *Philosophy Compass*, 3, 1–21.

Bearison, D. J. and Levey, L. M. (1977). Children's comprehension of referential communication: Decoding ambiguous messages. *Child Development*, 48, 716–20.

Beck, S. R. and Robinson, E. J. (2001). Children's ability to make tentative interpretations of ambiguous messages. *Journal of Experimental Child Psychology*, 79, 95–114.

Beck, S. R., Robinson, E. J., and Freeth, M. M. (2008). Can children resist making interpretations when uncertain? *Journal of Experimental Child Psychology*, 99, 252–70.

Call, J. and Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition*, 4, 207–20.

Carlson, S. M., Moses, L. J., and Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*, 11, 73–92.

Carlson, S. M., Moses, L. J., and Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology*, 87, 299–319.

Cepeda, N. J., Kramer, A. F., and Gonzalez de Sather, J. C. M. (2001). Changes in executive control across the life span: Examination of task-switching performance. *Developmental Psychology*, 37, 715–30.

David, N., Gawronski, A., Santos, N., *et al.* (2008). Dissociation between key processes of social cognition in autism: impaired mentalizing but intact sense of agency. *Journal of Autism and Developmental Disorders*, 38, 593–605.

Flavell, J. H., Speer, J. R., Green, F. L., August, D. L., and Whitehurst, G. J. (1981). The development of comprehension monitoring and knowledge about communication. *Monographs of the Society for Research in Child Development*, 46, 1–65.

Frye, D., Zelazo, P. D., and Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, 10, 483–527.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Science*, 4, 14–21.

Gallagher, S. (2010). Multiple aspects in the sense of agency. *New Ideas in Psychology*, 30(1), 1–13.

Gallagher, S. and Zahavi, D. (2008). *The phenomenological mind: An introduction to philosophy of mind and cognitive science*. London: Routledge.

Gopnik, A. and Graf, P. (1988). Knowing how you know: Children's understanding of the sources of their knowledge. *Child Development*, 59, 1366–71.

Hill, E. L. (2004). Executive dysfunction in autism. *Trends in Cognitive Sciences*, 8, 26–32.

Hughes, C. (1998). Finding your marbles: Does preschoolers' strategic behavior predict later understanding? *Developmental Psychology*, 34(6), 1326–39.

Ironsmith, M. and Whitehurst, G. J. (1978). The development of listener abilities in communication: How children deal with ambiguous information. *Child Development*, 49, 348–52.

Johnson, C. N. and Wellman, H. M. (1980). Children's developing understanding of mental verbs: Remember, know, and guess. *Child Development*, 51, 1095–102.

Kelley, C. M. and Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, 32, 1–24.

Kloo, D., Perner, J., and Giritzer, T. (2010). Object-based set-shifting in preschoolers: Relations to theory of mind. In J. Carpendale, G. Iarocci, U. Müller, B. Sokol, and A. Young (Eds.) *Self- and social-regulation: Exploring the relations between social interaction, social cognition, and the development of executive functions*, pp. 193–217. Oxford: Oxford University Press.

Kloo, D. and Pinnitsch, A. (2010). The sense of agency in preschool children. Unpublished raw data.

Kochanska, G., Murray, K., Jacques, T. Y., Koenig, A. L., and Vandegeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development*, 67, 490–507.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609–39.

Lang, B. and Perner, J. (2002). Understanding of intention and false belief and the development of self-control. *British Journal of Developmental Psychology*, 20, 67–76.

Mazzoni, G. and Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1263–74.

Metcalfe, J. and Greene, M. J. (2007). Metacognition of agency. *Journal of Experimental Psychology: General*, 136, 184–99.

Metcalfe, J., Eich, T. S., and Castel, A. (2010). Metacognition of agency across the lifespan. *Cognition*, 116, 267–82.

Miscione, J. L., Marvin, R. S., O'Brien, R. G., and Greenberg, M. T. (1978). A developmental study of preschool children's understanding of the words 'know' and 'guess'. *Child Development*, 49, 1107–13.

Moses, L. J. and Baird, J. A. (1999). Metacognition. In R. Wilson (Ed.) *Encyclopedia of cognitive neuroscience.* Cambridge, MA: MIT Press.

Nelson, T. O. and Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.) *The psychology of learning and motivation: Advancesin research and theory*, pp. 125–73. New York: Academic Press.

O'Neill, D. K., Astington, J. W., and Flavell, J. H. (1992). Young children's understanding of the role that sensory experiences play in knowledge acquisition. *Child Development*, 63, 474–90.

O'Neill, D. K. and Chong, C. F. (2001). Pre-school children's difficulty understanding the types of information obtained through the five senses. *Child Development*, 72, 803–15.

O'Neill, D. K. and Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology*, 27, 390–7.

Pacherie, E. (1997). Motor-images, self consciousness and autism. In J. Russell (Ed.) *Autism as an executive disorder*, pp. 215–55. Oxford: Oxford University Press.

Patterson, C. J., Cosgrove, J. M., and O'Brien, R. G. (1980). Non-verbal indicants of comprehension and noncomprehension in children. *Developmental Psychology*, 16, 38–48.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Perner, J. (1998). The meta-intentional nature of executive functions and theory of mind. In P. Carruthers and J. Boucher (Eds.) *Language and thought*, pp. 270–83. Cambridge: Cambridge University Press.

Perner, J. and Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences*, 3, 337–44.

Perner, J. and Ruffman, T. (1995). Episodic memory and autonoetic consciousness: Developmental evidence and a theory of childhood amnesia. *Journal of Experimental Child Psychology*, 59, 516–48.

Perner, J., Lang, B., and Kloo, D. (2002). Theory of mind and self-control: More than a common problem of inhibition. *Child Development*, 73, 752–67.

Pillow, B. H. (1989). Early understanding of perception as a source of knowledge. *Journal of Experimental Child Psychology*, 47, 116–29.

Pillow, B. H. (1993). Pre-school children's understanding of the relationship between modality of perceptual access and knowledge of perceptual properties. *British Journal of Developmental Psychology*, 11, 371–89.

Pillow, B. H. (2002). Children's and adults' evaluation of the certainty of deductive inferences, inductive inferences, and guesses. *Child Development*, 73, 779–92.

Phillips, W., Baron-Cohen, S., and Rutter, M. (1998). Understanding intention in normal development and in autism. *British Journal of Developmental Psychology*, 16, 337–48.

Plummert, J. M. (1996). Young children's ability to detect ambiguity in descriptions of location. *Cognitive Development*, 11, 375–96.

Pratt, C. and Bryant, P. (1990). Young children understand that looking leads to knowing so long as they are looking into a single barrel. *Child Development*, 61, 973–82.

Robinson, E. J. and Whittaker, S. J. (1985). Children's responses to ambiguous messages and their understanding of ambiguity. *Developmental Psychology*, 21, 446–54.

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., and Apperly, I. A. (2006). Children's sensitivity to their own relative ignorance: Handling of possibilities under epistemic and physical uncertainty. *Child Development*, 77, 1642–55.

Robinson, E. J., Haigh, S. N. and Pendle, J. E. C. (2008). Children's working understanding of the knowledge gained from seeing and feeling. *Developmental Science*, 11, 299–305.

Rohwer, M. A., Kloo, D., and Perner, J. (2010). Children fail in partial exposure task settings even when allowed to express their ignorance with non-verbal responses. Unpublished raw data.

Rohwer, M. A., Kloo, D., and Perner, J. (2012a). *Escape from meta-ignorance: How children develop an understanding of their own lack of knowledge*. Manuscript accepted for publication.

Rohwer, M. A., Kloo, D., and Perner, J. (2012b). *Understanding of own ignorance in children: Explicit acknowledgement versus implicit sensitivity*. Manuscript in preparation.

Ruffman, T. and Olson, D. R. (1989). Children's ascriptions of knowledge to others. *Developmental Psychology*, 25, 601–6.

Russell, J. (1996). *Agency. Its role in mental development*. Hove: Erlbaum.

Russell, J. and Jarrold, C. (1998). Error-correction problems in autism: evidence for a monitoring impairment? *Journal of Autism and Developmental Disorders*, 28, 177–88.

Schult, C. A. (2002). Children's understanding of the distinction between intentions and desires. *Child Development*, 73, 1727–47.

Shultz, T. R., Wells, D., and Sarda, M. (1980). The development of the ability to distinguish intended actions from mistakes, reflexes and passive movements. *British Journal of Social and Clinical Psychology*, 19, 301–10.

Sekerina, I., Stromswold, K., and Hestvik, A. (2004). How do adults and children process referentially ambiguous pronouns? *Journal of Child Language*, 31, 123–52.

Sodian, B. and Wimmer, H. (1987). Children's understanding of inference as source of knowledge. *Child Development*, 58, 424–33.

Somerville, S. C., Hadkinson, B. A., and Greenberg, C. (1979). Two levels of inferential behavior in young children. *Child Development*, 50, 119–31.

Tardif, T., Wellman, H. M., Fung, K. Y. F., Liu, D., and Fang, F. (2005). Preschoolers' understanding of knowing- that and knowing- how in the United States and Hong Kong. *Developmental Psychology*, 41, 562–73.

Taylor, M., Esbensen, B. M., and Bennet, R. T. (1994). Children's understanding of knowledge acquisition: The tendency for children to report to have always known what they have just learned. *Child Development*, 65, 1581–604.

Wimmer, H. (1989). Common-sense Mentalismus und Emotion: Entwicklungspsychologische Implikationen. I. In E. Roth (Ed.), *Denken und Fühlen*, pp. 56–66. Berlin: Springer.

Wimmer, H., Hogrefe, G. J., and Perner, J. (1988). Children's understanding of informational access as a source of knowledge. *Child Development*, 59, 386–96.

Zakay, D. and Tuvia, R. (1998). Choice latency times as determinants of post-decisional confidence. *Acta Psychologica*, 98, 103–15.

Zelazo, P. D. (2006). The dimensional change card sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, 1, 297–301.

Zelazo, P. D., Frye, D., and Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development*, 11, 37–63.

Chapter 11

Thinking about different types of uncertainty

S. R. Beck, E. J. Robinson, and M. G. Rowley

Imagine you are searching for your lost keys. Someone helpfully asks, 'Where did you last see them?' to which you reply that you don't know. Your response could be based on the fact that you simply have no idea of their location, but you may recognize that there is a set of possible locations where they may be. Certainly to resolve your problem and find the keys you need to identify these possible locations. Identifying possibilities can directly inform a reflective metacognitive state: not knowing where the keys are but being pretty sure you left them either in your office or in the lecture theatre is rather different from drawing a complete blank on where they might be, but there are also other ways in which thinking about possibilities is relevant for our understanding of metacognition.

We know that children have some difficulty understanding metacognition. Are they able to handle uncertainty and possibilities at a younger age, before they develop this full metacognitive understanding? If so, what does this tell us about what is difficult about metacognition? Perner's minimeta system (Chapter 6, this volume) suggests there are different levels of cognition that underpin full blown metacognitive understanding. Being able to recognize that one is uncertain and identify different types of possibilities is not the same as full metacognitive understanding, which we take to involve a reflective evaluation of one's epistemic state (what Perner would call recursive cognition). Yet, it seems likely that handling uncertainty and being able to identify possibilities is necessary to make such metacognitive judgements. Being able to identify two different possible worlds (the keys are in your office or they are in the lecture theatre) requires generating two alternative models of how the world might be. Perner sees generating alternative models as a part of special cognition or thinking '*beyond* object cognition' and as such it forms part of the minimeta system that is 'on the way to metacognition' (Chapter 6, this volume). Our research focuses primarily on this ability to generate alternative models of how the world might be.

In this chapter we review recent work on children's handling of uncertainty, and we make links to our own and others' work with adults. In doing so we identify differences in how children deal with various types of uncertainty. These different types have been neglected in the developmental literature, yet, understanding children's responses to them will shed light on the development of children's handling of uncertainty and hence their developing metacognition. To a large extent the developmental literature has focused on children's difficulties with explicit reflective metacognitive judgements (e.g. Ironsmith and Whitehurst 1978; Beck and Robinson 2001) and the apparently appropriate behaviour that accompanies this (eye movements, hesitation, etc.; e.g. Plumert 1996), with an ensuing debate as to whether these non-verbal behaviours reflect implicit 'understanding' of uncertainty or something more basic such as a switching between two possibilities without simultaneously holding both in mind (see Beck et al. 2008). Yet the claims based on reflective judgements and non-verbal behaviours remain at a stalemate. A more careful examination

of children's handling of different types of uncertainty should advance our understanding of developing metacognition. Firstly, it might be that exploring different types of uncertainty will reveal full-blown metacognition earlier than we expected. Secondly, we may find evidence for developmental precursors to full recursive cognition (Perner's minimeta stages, or what Esken describes as the 'first hint concerning the ontogenetic development of these abilities' (Chapter 8, this volume)).

Can children acknowledge possibilities?

Our first question was whether children are able to mark the multiple possibilities that arise under conditions of uncertainty. Sophian and Somerville (1988) used a task in which a toy was hidden in one of several cups and the child had to place a mat underneath it to ensure that it was caught when it was tipped from its hiding place. Sometimes children could know for certain where the toy was hidden, but sometimes there were multiple possible hiding places. In one sense performance was relatively good in that even 4-year-olds were sensitive as a group to the situation and tended to put out more mats when they could not be sure where the toy was. On the other hand, 6-year-olds were rarely consistent in their responses, with only four individuals from a sample of 16 reliably covering the right possibilities on uncertain problems.

We (Beck et al. 2006) designed a very simple procedure to investigate this further and in doing so we uncovered a distinction between different types of uncertainty that had been overlooked by the developmental literature, although it had received some attention in the adult cognitive literature (e.g. Heath and Tversky 1991; and see more recently Fox and Ülkümen 2011). In two experiments, children (aged 3–6 years) saw a toy mouse run down one of two slides. The red slide had an inverted Y shape resulting in two possible exits. The blue slide was straight, so there was only one exit. On each trial, once children knew which slide the mouse was going to take they were instructed to 'put out cotton wool to make sure he lands safely'. The children had already been familiarized with a set of cotton wool mats, learnt the importance of making sure the mouse did not crash land on the floor, and practised putting out mats. Even the youngest 3- to 4-year-old group were more likely to put out two mats when the mouse was about to come down the red slide (21% of trials) compared to the blue (1%). Yet their performance was relatively poor, as they were much more likely to place only one mat than two on the red slide trials. Even more compelling is the fact that children continued to avoid putting out two mats when after they had placed just one the experimenter prompted them, 'Could it [i.e. the mouse] go anywhere else?'. On hearing this most of the younger children either refused or moved their single mat to the other exit. On only 10% of trials did this prompt lead to the youngest children adding an extra mat to hedge their bets. Overall, performance on this task improved substantially between 3–6 years of age. The very youngest children tested, who had a mean age of 3 years and 7 months, put out two mats on 30% of the red trials (even with the benefit of the prompt). Four- to 5-year-olds placed two mats on 60% of trials and 5- to 6-year-olds did so on 85%.

The youngest children struggled somewhat with the task, but the older children did quite a good job of putting out two mats to ensure the mouse was caught. In other words, they were able to mark both possible future outcomes. Performance looked good compared to Sophian and Somerville's (1988) study and we speculated that one reason for this was that the type of uncertainty in the two studies differed.

In Sophian and Somerville's study the toy was hidden in one of the locations at the point when the child had to mark the multiple possibilities. However, in the Beck et al. study the mouse had yet to run down the slide. Throughout this chapter we will use 'epistemic uncertainty' to describe the first situation: where there is a fact of the matter but it is unknown to the individual acting on it, and 'physical uncertainty' to describe the second: there is as yet no fact of the matter and the outcome is undetermined. Was it possible that children found it much easier to handle the

physical uncertainty involved in the mouse game (Beck et al. 2006) than the epistemic uncertainty that had dominated the established literature? To investigate this, Robinson et al. (2006) made direct comparisons between trials involving epistemic and physical uncertainty. In their first experiment they used an apparatus called the Doors game. The Doors game consisted of a cardboard screen coloured in three stripes with three doors in it. Small blocks in the same colours as the stripes and doors could be placed on a shelf behind the doors and were pushed through the doors by the experimenter. The child's job was to catch the blocks, by placing trays under the doors where the blocks might fall. Blocks were stored in two bags. One contained only black blocks and the other contained a mixture of green and orange blocks. Blocks always fell through their matching coloured door, which meant that if you knew the block came from the black bag you could be sure which door it would emerge from. However, if the block came from the orange and green bag there were two possible doors through which it could fall. Thus, there were two possible responses children should make. If the block was picked from the black bag, then children only needed to put a tray under the black door. This is what the majority of 4- to 6-year-olds did. However, if the orange and green bag were used then they needed to hedge their bets and place trays under both the orange and the green doors. To compare handling of epistemic and physical uncertainty, Robinson et al. used a simple manipulation. In physical trials children placed trays (or a tray) before the block had been picked from the orange and green bag. On epistemic trials the block was in place behind the matching coloured door before they placed the trays, but children did not know which colour it was (nor which door it was behind), only that it had come from the orange and green bag. This simple manipulation had a clear effect on children's performance. Children put out two trays, thus ensuring the block was caught, on 75% of physical trials (where the block had yet to be picked) but on only 41% of epistemic trials (where the block was in place, waiting to fall). They did so on only 18% of trials where they knew that the block was black, as it had come from the black bag.[1]

In two further experiments, Robinson and colleagues confirmed this finding using a rather different type of task, the Pet Shop game. Children heard a story about a pet shop owner who needed to transport pets in boxes that varied in size and colour. In epistemic trials they were given a message that a pet was already in a box, but the message could refer to more than one box. For example 'The mouse is in the large box' when there were in fact two large boxes. Children had to make sure that the pet was fed for his journey and so when there were two possible boxes that could contain the pet, the correct response was to place food in both large boxes. In physical trials the pet had not yet arrived, but the child was told that someone would be bringing a pet to go in one of the boxes. The child had to prepare the box(es) by pressing switches to heat them (each box had its own switch). So, if the message indicated that the pet might need to go in either of two boxes, the correct response was to heat both just in case. Only 5- to 6-year-olds participated in this game, but performance was very similar to that in the Doors game experiment: children prepared two boxes on 75% of physical trials, but put food in two boxes on only 40% of epistemic trials. Poor performance on epistemic trials was replicated in another experiment using the Pet Shop game. Furthermore, a mental state measure was included in this final experiment. Children had to choose which of three thought bubbles best represented what the pet shop owner was thinking. The thought bubbles contained (1) just a white large box, (2) just a pink large box, (3) both large boxes with the word 'OR' between them and a question mark. Although 5- to 6-year-olds found the thought bubble task difficult in both conditions, 7- to 8-year-olds showed

[1] These 18% of participants were equally likely to place one or two mats on epistemic trials suggesting that children's performance may have been worse than first appears as some children may not have had a firm grasp of when it was appropriate to place one or two mats. The difference between the epistemic and physical trials remained when children who placed two mats on the black trial were excluded.

the same pattern of relative difficulty as the younger children had in the behavioural task: they were more likely to pick the thought bubble with both boxes in the physical condition (80% correct) than the epistemic condition (45%). That the 7- to 8-year-olds seem to be showing the same difficulty handling multiple possibilities under epistemic uncertainty that we may have thought the 6-year-olds had overcome based on the Doors game might suggest that this understanding is rather fragile and the extra theory of mind demands of the thought bubble task (thinking about other people, for example) might reduce the resources the child has to devote to handling the possibilities. Alternatively, we will see in a later section that a bias in handling epistemic and physical uncertainty is preserved even into adulthood, although of course they would not make the error that the 7- to 8-year-olds make here. Perhaps the 7- to 8-year-olds' behaviour reflects something like the bias seen in adults.

Differentiation between epistemic and physical uncertainty has been shown in another procedure, where children also had to mark different possibilities, but there was more explicit reference to making a guess. In this game there were three boxes each containing a different toy. One box was (or was to be) picked from this set and before its contents were revealed children had to make a guess about which toy it contained by placing a marker on one of three corresponding pictures. They were then given the opportunity to improve their chance of correctly guessing the identity of the toy by making a second guess on behalf of another person (who they also wanted to help guess correctly). Five- to 7-year-olds showed a tendency to make the same guess for themselves and the other, but they made different guesses more often when the uncertainty was physical (the box had not yet been picked from the set of three) than when it was epistemic (the box had been picked): 41% compared to 27% respectively (McColgan et al. unpublished manuscript).

Overall, in several different procedures we found that children were consistently more likely to acknowledge possibilities under physical rather than epistemic uncertainty. This was the case whether they were asked to make a behavioural response on their own or someone else's behalf or whether they had to report what someone else would think. This different treatment of physical and epistemic uncertainty had not been investigated previously by the developmental literature, resulting in an overly pessimistic view about children's ability to acknowledge possibilities under uncertainty. Children's relative difficulty acknowledging possibilities under epistemic uncertainty may mean that thinking about knowledge under these conditions makes extra demands on children over just acknowledging their ignorance under physical uncertainty. For example, appreciating epistemic uncertainty may involve representing what one knows and what one could know (see Esken's discussion of evaluative emotions, Chapter 8, this volume). Thus, handling epistemic uncertainty may make further demands at the truly metacognitive (recursive) level. Alternatively, the difference may be at the level of special cognition (Perner Chapter 6, this volume). Perhaps it is easier to create alternative models of possible outcomes under physical uncertainty rather than epistemic. A third possibility is that when children appreciate physical uncertainty they can also reflect on this uncertainty metacognitively. We return to these possibilities having considered further evidence on children's handling of these types of uncertainty.

What does it mean for the child's experience to be able to acknowledge possibilities? In our next section we review work where we explored whether children behaved as if they were less confident when they had identified multiple possibilities.

Preferences for guessing under epistemic and physical uncertainty

We decided to examine children's relative confidence under epistemic and physical uncertainty. We adopted a strategy through which we could infer whether children felt more confident under

one type of uncertainty compared to another. In this paradigm, children played two versions of a simple chance game: one involving epistemic uncertainty and one involving physical. They were then given a choice of how to play the game for a third iteration in which a sticker was offered as an incentive. We assumed that a preference to guess under one type of uncertainty indicated greater confidence in being able to predict the outcome. However, as we will discuss, there remain questions to be addressed about what behavioural responses and direct evaluations of confidence indicate about children's metacognition.

In our first experiments on children's guessing preferences, we used a Die game (Robinson et al. 2009). Our first sample comprised 5- to 6-year-olds and 7- to 8-year-olds. Each child practised throwing the die under a cup and then played two practice guessing trials. On one trial the child guessed what number would come up on the die and then threw it and on the other trial s/he threw the die under the cup first and then guessed what number had come up. The order of these two practice trials was counterbalanced. For the third trial the experimenter explained that if the child guessed correctly she would win a sticker. He then asked, 'So what do you want to do? Guess first then shake or shake first then guess?'. We analysed the results in two groups. First we considered only children who had been successful or unsuccessful on both of the practice trials, i.e. success on one version of the task could not have influenced their choice. Nineteen of 20 5- to 6-year-olds and 17 of 24 7- to 8-year-olds chose to guess[2] after the die had been thrown. Even when we included the children who had been lucky on one practice trial, the preference for guessing after the event, i.e. under epistemic uncertainty, remained.

In a follow-up experiment we replicated this finding (McColgan et al. unpublished manuscript). We used two new procedures with children aged 5–6 years old. One was a modified version of the Die game in which once the die had been thrown under a cup the experimenter moved the die behind a screen on which were depicted the six faces of a standard die. Children used a pointer to indicate which number they thought had or would come up on the die. The other game was the Doors game, which we had used previously with the acknowledging possibilities measure. In the preference version children played with only one mat. In both games children had a practice trial where they guessed under physical uncertainty (before the die had been thrown, or the object picked in the doors game) and a practice trial where they guess under epistemic uncertainty (the die or the object was in place behind the screen or the door). Our findings were clear. In both games children showed a significant preference to guess under epistemic rather than physical uncertainty. Twenty-one children (out of 28) did so in the Die preference game and 20 did so in the Doors game. Note that the difference in number of possibilities (six in the Die game and three in the Doors game) had no bearing on children's preference for guessing under epistemic uncertainty.

Finally, we used a rather different narrative task to investigate children's preference for guessing under epistemic or physical uncertainty. Five- to 6-year-olds were shown a toy farm where three animals—a cow, a sheep and a pig—stood in a field with Mr Farmer and a truck. Mrs Farmer stood by a barn that was reached from the field by a road. A cardboard 'spinner' (an arrow pinned to a cardboard base which when spun landed on a picture of one of the animals) was used to determine which of the three animals was sent in the truck to the farm where it was fed by Mrs Farmer. The animal was placed in the truck behind a screen and the truck was covered so the child could not see which animal had been chosen. Children practised guessing which animal was

[2] Note that we call this behaviour 'guessing' which reflects the objective state of the world. The child does not know the outcome e.g. of the die throw, and so can be described as guessing it. However, it is possible that the child is inappropriately confident in her answer. In this case, from the child's point of view the response is not a guess.

going to the barn on two practice trials: before the spinner had been spun and after the truck had reached the barn (but before it was opened). The third trial, as before, offered them a choice of when to guess. As in the other preference games, 5- to 6-year-olds showed a just significant preference ($p = 0.049$) to guess under epistemic uncertainty (when the truck was at the barn). Thirteen of 17 children chose this version of the game for their third trial.

Thus, children's ability to acknowledge possibilities appears to translate into some sense of confidence: they showed a consistent preference to guess under epistemic uncertainty compared with physical. In the next studies, we explored children's explicit knowledge evaluations to see if there was good evidence that children were demonstrating recursive cognition and recognizing that they were less confident under physical uncertainty.

Evaluations of knowledge

We inferred from their guessing preferences that children feel more confident when uncertainty is epistemic rather than physical. But we did not yet know whether this behaviour reflected full-blown recursive metacognitive understanding. Unfortunately there is very little direct evidence to address this. However, we will summarize the data we have collected and identify some of the questions that remain to be addressed.

To investigate explicit confidence judgements we used a version of the Doors game adapted from Robinson et al. (2006). There were just two doors and children were given one mat that they used to guess where the block would fall from. Having placed the mat they then rated how confident they felt about their guess using a five-point scale (adapted from Pillow and Anderson 2006). Children played the game under two conditions: in the epistemic condition they placed the mat and rated their confidence once the block was in place behind the door, in the physical condition they placed and rated before the block was in place. There was no difference in children's confidence ratings between the trials. In fact, children tended to report that they were very confident in both conditions.

Children's explicit evaluations of their certainty were in keeping with findings from the established literature that they tend to be overconfident in the face of uncertainty (e.g. Ironsmith and Whitehurst 1978; Beck and Robinson 2001). Yet this overconfidence was seen when there was physical as well as epistemic uncertainty. We found this surprising and made numerous other attempts to investigate children's confidence ratings under physical uncertainty, but to date have found no reliable evidence that children are any less confident under physical uncertainty than epistemic. This is in clear contrast to their behavioural responses to epistemic and physical uncertainty. Children were more likely to identify only one possible outcome in tasks where they had to acknowledge possibilities (e.g. put out mat(s) to catch a falling block) and in tasks where they had a choice of whether to guess under physical or epistemic uncertainty they preferred to guess under the latter (e.g. guess after the die has been thrown rather than before). We had inferred from this behaviour that children experienced greater confidence under epistemic uncertainty. However, the lack of a difference in confidence ratings between the two types of uncertainty suggests that children's ability to acknowledge possibilities can develop in advance of an ability to evaluate one's confidence. It remains possible that changes to our methodology may reveal differences in confidence ratings. For example, we could use a more sensitive scale or asking children to rate their confidence before rather than after they had made an overt guess. However, we have no reason to think that these adaptations would specifically affect the physical uncertainty trials and so at the current time we conclude that children's explicit evaluations do not differentiate these types of uncertainty, despite the differences in children's behaviour.

Answering explicit confidence evaluation questions appropriately requires recursive metacognitive understanding. We saw overconfidence on both epistemic and physical trials. Thus, we

have no evidence to think that children are showing recursive or full metalevel cognition when they play our games, even under physical uncertainty.

That children prefer not to guess under epistemic uncertainty compared to physical and find it more difficult to mark multiple possibilities under the former may be due to the fact that handling epistemic uncertainty necessarily makes metacognitive demands (that children at this age cannot meet), i.e. that you have to evaluate what you know compared to what could be known. On the other hand, it may be easier for children to build alternative models of the world under physical rather than epistemic uncertainty. There may be reasons other than recursive metacognitive demands why building alternative models is more difficult under epistemic uncertainty compared to physical. In the next section we consider this possibility.

Mechanisms for handling uncertainty

There is now a body of evidence that children treat epistemic and physical uncertainty differently in their behaviour (even though their explicit evaluations of knowledge do not appear to be influenced by this differentiation). Children prefer to make guesses under epistemic rather than physical uncertainty and, in other tasks, they are less likely to acknowledge multiple possibilities when uncertainty is epistemic rather than physical. They behave *as if* they are overconfident. But we are left without explanation of why this is the case. We turned to this question of why children treat epistemic and physical uncertainty differently in our next experiments.

We considered the possibility that once the outcome has occurred children might imagine one version of this outcome. We know that children have rich imaginative abilities (e.g. Harris 2000). Perhaps the ease with which they can imagine one possible outcome results in a sense of confidence that this is the actual outcome. This account is closely related to fluency effects that we see in adults (e.g. Alter and Oppenheimer 2009). In order to test between these accounts we manipulated how easy it was for children to imagine the outcome under epistemic uncertainty.

In our first experiment, children played a version of the Doors game in which the door that the object would be placed behind was determined by the throw of a die with faces of three different colours (Beck et al. 2011). We used the 'Specified' condition in which children knew the object behind the door was a yellow pom pom and an 'Unspecified' condition in which children did not know the identity of the 'something' behind the door. We reasoned that knowing what the object was made it easier to imagine in place behind the door. So although the location was equally equivalent in both versions of the game the likelihood of imagining the object in place was not. If the ease with which children can imagine an outcome is relevant to their handling of uncertainty, then we would expect children's performance on the two conditions to reflect this. In each condition, children practised a physical and an epistemic version of the game and then chose which way to play on a third trial. Performance by a group of 5- to 6-year-olds in the Specified condition was similar to that in previous studies: 45 of 61 children chose to guess after the pom pom was in place. However, when children did not know what the object was that was hidden behind the door 32 children chose to guess once it was in place but 29 chose to guess before under physical uncertainty. There was no preference for one version of the game over the other.

In a second experiment we examined whether the specified manipulation would influence children's marking of possibilities. In this study we used only epistemic trials, half of which were specified (the child knew it was the yellow pom pom hidden behind one of the doors) and half were unspecified ('something' was hidden). The appropriate cautious response was to place two mats on all trials to cover the two doors from which the object could fall. Children's performance was much better when they did not know what the object was behind the door (16 of 29 children always placed two mats on Unspecified trials), compared to when they did know (8 of 29 children always placed two mats on Specified trials).

Our results were in line with the idea that children are imagining a possible outcome. If their preference resulted from them imagining an outcome under epistemic uncertainty then we would expect any preference for guessing under these conditions to disappear when they did not know what the object was. This is what we observed. In the preferences experiment children's preference to guess under epistemic uncertainty disappeared when they did not know the identity of the hidden object. Furthermore, in the possibilities experiment, when children did not know the identity of the object they were relatively good at marking the multiple possible locations where it could be. We concluded that one reason children have difficulty handling epistemic uncertainty is that they imagine one possible outcome (see also Kloo and Rohwer (Chapter 10, this volume), who found that when children could bring a possible answer to mind they were likely to overestimate their knowledge about it). When this is difficult to do (e.g. in the Unspecified condition they are better able to handle the uncertainty) we speculate that one reason children are more likely to mark multiple possibilities and avoid guessing under physical uncertainty is that they are less likely to imagine an outcome that has yet to happen. Being able to imagine a possible outcome might impair children's ability to make metacognitive evaluations. This would mean that an important development in children's metacognition might be to recognize imagined outcomes for what they are, namely uncertain possibilities.

We suggest that these findings support the second possibility that we considered at the end of the last section: it is easier for children to build alternative models under physical rather than epistemic uncertainty. However, the first possibility: that handling epistemic uncertainty necessarily makes inherent metacognitive demands would be a worthwhile avenue for further research.

Adults' behaviour and unanswered developmental questions

One question we have not yet addressed is whether adults are also influenced by the difference between epistemic and physical uncertainty. This seems to be the case. Until recently it was thought that adults behaved as if they were less confident under epistemic uncertainty. They prefer to guess under physical uncertainty, place higher bets under these conditions, and give higher certainty ratings (Rothbart and Snyder 1970; Brun and Teigen 1990; Chow and Sarin 2002; see also Robinson et al. 2009, experiment 1). Heath and Tversky (1991) offered the competence account as an explanation for these behaviours. Adults are averse to feelings of incompetence and these are more likely to arise when there is something that could be known (epistemic uncertainty) than when there is no unknown fact of the matter (physical uncertainty): they make a counterfactual comparison between what they know and what they could know.

These claims were based largely on studies where participants had to simulate the different situations, e.g. they had to imagine a game of chance or an imaginary stock market. Our developmental work led us to test adults on live versions of the tasks, producing a surprising result. When adults played simple chance games live (e.g. guessing what number would come up on a die) their preference for physical uncertainty was reversed. In the live version of the game they preferred to guess under epistemic uncertainty just like young children (Robinson et al. 2009, experiment 4) and their performance was significantly different under live and imagined versions of the game (Robinson et al. 2009, experiment 3). We have no explanation as yet about why adults' behaviour should differ when they imagine the game or play it live and this question remains to be addressed. Another question is whether children also treat simulated uncertainty differently to real-life experiences. We have found in preliminary work with adolescents that the preference for physical uncertainty under imagined conditions seems to emerge around 15 years. But we have yet to identify what developmental process may underpin this shift.

Despite these questions, the competence account highlights a factor that may influence adults' and children's handling of uncertainty. Adults appear to make comparisons between what they know and what could be known. Other research with adults suggests that one situation where this is pertinent is when one is ignorant while others are knowledgeable. Chow and Sarin (2002) found that adults bet more in situations where no one knew the outcome of an uncertain event, compared to situations where the outcome was known to others but not to them. This comparison with what other people (may) know appears to be relatively late developing. We played a preference version of the die game with 5- to 8-year old children. In all cases the die was thrown under the cup when the child guessed, but we manipulated whether no one knew what the outcome was, or whether the experimenter knew (she had peeked under the cup). We found no evidence that children preferred situations when their own ignorance was matched by the other's.

One possibility is that for the difference between self and other's knowledge to be salient one needs to spontaneously compare what is known with what could be known (as represented by the other person). This would involve counterfactual thinking. We know that children can pass some simple counterfactual conditional tasks at around 4 years (e.g. Riggs et al. 1998; Guajardo and Turley-Ames 2004) and this speculative thinking continues to develop through middle childhood (Beck et al. 2006; Beck and Guthrie 2011) with some authors suggesting that the critical distinctions between counterfactual and hypothetical worlds are not appreciated until children are around 10 years old (Rafetseder and Perner in press). One important unanswered question in the developmental literature is when children make spontaneous counterfactual comparisons compared to those prompted by the experimenter. Perhaps if children's attention were drawn to the counterfactual they may be influenced by other people's knowledge in our uncertainty tasks. Understanding when children make comparisons (either prompted or spontaneously) with what other people know will inform both the counterfactual and theory of mind literatures.

The influence of others' knowledge on adults' handling of uncertainty illustrates that adults are influenced by 'social' or 'mental' factors when they evaluate uncertainty in the world. In recent work we have investigated another social factor that affects adults' treatment of uncertainty: agency. A vast literature on the Illusion of Control (originating from Langer 1975) shows that adults prefer to guess about chance events which are to some extent under their control rather than those which are controlled by others or physical means. We also know that adults are more likely to generate counterfactual thoughts (speculations about what might have been) when a past event was under their control, than when it was not (e.g. Zeelenberg et al. 1998) and that children's counterfactual thinking is influenced by this difference (Weisberg and Beck in press).

To explore this we devised a new game to use with the preference paradigm (Harris et al. 2011). In the Pens game five identical looking pens were held in a pot. Despite being indistinguishable when capped, they each had different coloured ink. In the game one pen was picked from the pot and used to draw a circle on a piece of paper that was hidden from sight. The participant had to guess the colour of the circle or the colour of the pen either before the pen was picked (physical) or after the circle was drawn (epistemic), NB whether they were asked about the circle or the pen made no difference to the results. Remember that in both epistemic and physical conditions the drawing took place out of sight, behind a screen, so the participant could not see the colour of the circle or the pen. The critical difference was whether the participant selected the pen and drew the circle (thus having some sense of control over the outcome) or the experimenter did so. Adults treated these two conditions differently. Just like in our earlier preference game with children (e.g. Robinson et al. 2009, described in the section 'Preferences for guessing under epistemic and physical uncertainty') participants played two practice versions of the game: one physical and one epistemic. They were then given a choice of which way to play the game for a third trial.

When participants were the ones who picked the pen (not the experimenter), they were more likely to choose to play the physical version on the third trial. In fact, in line with the competence hypothesis they were significantly more likely to choose this than epistemic uncertainty (79 chose physical and 45 chose epistemic). When the experimenter controlled the pen there was no significant preference for guessing under either condition, although more participants chose to guess under epistemic uncertainty (38) compared to physical (27). The proportions of participants choosing to guess under epistemic uncertainty were somewhat higher in the Die game experiments reported in Robinson et al. (2009) and differences in the strength of adults' preferences in different procedures warrants further research. Indeed, understanding the mechanisms that underlie adults' performance in general is a rich area for further investigation. Are adults' biases based on cognitive demands (e.g. additional counterfactual thinking demands or fluency effects) or are they influenced by social factors (e.g. in an ambiguity avoidance task participants appeared to behave as if they were competing with another player who may try to prevent their success; Kühberger and Perner 2003)? Recent theoretical analysis of adults' behaviour (Fox and Ülkümen 2011) indicates that there is much more to be done to understand how adults think about different types of uncertainty.

In the Pens game it mattered to adults whether or not they were in control of the chance event. Even though the picking of the pen was uninformed and there was no real control over what colour the circle was, it seems that the act of 'choosing' a pen gave our participants an illusion of control (Langer 1975). This sense of control may have encouraged them to make counterfactual comparisons between what they know and what they could know, resulting in feelings of relative incompetence in the epistemic condition: I could know the outcome but I don't (the same relative incompetence is not possible under physical uncertainty because no one could know the outcome yet). Thus, given a sense of control, adults' judgements were in line with Heath and Tversky's (1991) competence account.

We played the same game with a group of 5- to 6-year-olds. Children in both conditions preferred to guess under epistemic uncertainty than physical (16 of 23 children in the child picks condition $p = 0.093$, 24 of 29 children in the experimenter picks $p = 0.001$). But they were not affected by the manipulation and showed the same pattern of behaviour in both child picks and experimenter picks conditions. Children were not susceptible to the same sense of whether they had 'controlled' the outcome that adults were. It remains a question for future research when children develop the sensitivity to agency that influences adults' thinking.

Conclusions

In this chapter we have reviewed new evidence on children's handling of uncertainty, specifically recognizing possibilities as a precursor to a full metacognitive ability to reflect on one's epistemic state. Understanding children's developing abilities to recognize possibilities firstly show us what children can do before they are able to make explicit metacognitive evaluations. Furthermore, by investigating possibilities we have identified hitherto neglected distinctions between different types of uncertainty which need to be fully investigated in terms of their influence on children's and adults' handling of uncertainty and metacognitive judgements. For example, children's performance on our tasks under conditions of physical uncertainty shows that they are able to acknowledge multiple possibilities rather earlier than previously thought. Children were able to mark two possibilities arising from physical uncertainty by 5 years of age, with some success even earlier (even 4- to 5-year-olds discriminated between epistemic and physical trials in Robinson et al.'s Doors game). Our results may be interpreted as evidence of minimeta cognition (Perner Chapter 6, this volume): children are able to represent alternative worlds, although only under some circumstances.

Both children and adults are influenced by whether uncertainty is epistemic or physical, whereas only adults are affected by others' minds when they deal with uncertainty (both in terms of what other people know and how events are determined). As yet, we have only shown the influence of these factors on participants' behaviour. It remains to be seen whether the same factors influence metacognitive judgements in development and adult cognition.

References

Alter, A. L. and Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13, 219–35.

Beck, S. R. and Guthrie, C. (2011). Almost thinking counterfactually: children's understanding of close counterfactuals. *Child Development*, 82, 1189–98.

Beck, S. R. and Robinson, E. J. (2001). Children's ability to make tentative interpretations of ambiguous messages. *Journal of Experimental Child Psychology* 79, 95–114.

Beck, S. R., Robinson, E. J., Carroll, D. J., and Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, 77, 413–26.

Beck, S. R., Robinson, E. J., and Freeth, M. M. (2008). Can children resist making interpretations when uncertain? *Journal of Experimental Child Psychology*, 99(4), 252–70.

Beck, S. R., McColgan, K. L. T., Rowley, M. G., and Robinson, E. J. (2011). Imagining what might be: why children under-estimate uncertainty. *Journal of Experimental Child Psychology*, 110, 603–10.

Brun, W. and Teigen, K. H. (1990). Prediction and postdiction preferences in guessing. *Journal of Behavioral Decision Making*, 3, 17–28.

Chow, C. C. and Sarin, R. K. (2002). Known, unknown and unknowable uncertainties. *Theory and Decision*, 52, 127–38.

Fox, C. R. and Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. B. Keren, G. Kirkebøen, and H. Montgomery (Eds.) *Perspectives on Thinking, Judging, and Decision Making*, pp. 21–35. Oslo: Universitetsforlaget.

Guajardo, N. R. and Turley-Ames, K. J. (2004). Preschoolers' generation of different types of counterfactual statements and theory of mind understanding. *Cognitive Development*, 19, 53–80.

Harris, A. J. L., Rowley, M. G., Beck, S. R., Robinson, E. J., and McColgan, K. L. T. (2011). Agency affects adults', but not children's, guessing preferences in a game of chance. *Quarterly Journal of Experimental Psychology*, 64, 1772–87.

Harris, P. L. (2000). *The work of the imagination*. Oxford: Blackwell.

Heath, C. and Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4, 5–28.

Ironsmith, M. and Whitehurst, G. J. (1978). The development of listener abilities in communication: How children deal with ambiguous information. *Child Development*, 49, 348–52.

Kühberger, A. and Perner, J. (2003). The role of competition and knowledge in the Ellsberg task. *Journal of Behavioral and Decision-Making*, 16, 181–91.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32, 311–28.

McColgan, K. L. T., Robinson, E. J., Beck, S. R., and Rowley, M. G. (Unpublished manuscript). Thinking about possibilities: knowing there is an unknown reality makes it harder.

Pillow, B. H. and Anderson, K. L. (2006). Children's awareness of their own certainty and understanding of deduction and guessing. *British Journal of Developmental Psychology*, 24, 823–49.

Plumert, J. M. (1996). Young children's ability to detect ambiguity in descriptions of location. *Cognitive Development*, 11, 375–96.

Rafetseder, E. and Perner, J. (in press). When the alternative had been better: counterfactual reasoning and the emergence of regret. *Cognition & Emotion*.

Riggs, K. J., Peterson, D. M., Robinson, E. J., and Mitchell, P. (1998). Are errors in false belied tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, 12, 73–90.

Robinson, E. J., Pendle, J., Rowley, M. G., Beck, S. R., and McColgan, K. L. T. (2009). Guessing imagined and live chance events: Adults behave like children with live events. *British Journal of Psychology*, 100, 645–59.

Robinson, E. J., Rowley, M., Beck, S. R., Carroll, D. J., and Apperly, I. A. (2006). Children's sensitivity to their own relative ignorance: handling of possibilities under epistemic and physical uncertainty. *Child Development*, 77, 1642–55.

Rothbart, M. and Snyder M. (1970). Confidence in the prediction and postdiction of an uncertain outcome. *Canadian Journal of Behavioral Science*, 2, 38–43.

Sophian, C. and Somerville, S. C. (1988). Early developments in logical reasoning: considering alternative possibilities. *Cognitive Development*, 3, 183–222.

Weisberg, D. P. and Beck, S. R. (in press). The development of children's regret and relief. *Cognition and Emotion*.

Zeelenberg, M., van Dijk, W. W., and Manstead, A. S. R. (1998). Reconsidering the relation between regret and responsibility. *Organizational Behaviour and Human Decision Processes*, 74, 254–72.

Chapter 12

Credulity and the development of selective trust in early childhood

Paul L. Harris, Kathleen H. Corriveau, Elisabeth S. Pasquini, Melissa Koenig, Maria Fusaro, and Fabrice Clément

Many recent studies have underlined the fact that, under certain conditions, 3- and 4-year-old children will defer to proposals that run counter to their own ideas and observations. In opening a box, they set aside their own efficient procedure to reproduce a more elaborate and inefficient technique that has been demonstrated to them (Horner and Whiten 2005; Lyons et al. 2007, 2011; Nielsen and Tomaselli 2010). Asked to say what category an object belongs to and infer its properties, they revise their initial, appearance-based categorization when an adult proposes an alternative that is less consistent with the available perceptual evidence (Jaswal 2004). When told about the movement and final resting-place of an object falling down an opaque tube, they are prepared to set aside their otherwise robust, gravity-based expectations to search where told (Bascandziev and Harris 2010; Jaswal 2010). Indeed, even when confronted with repeated evidence that what they have been told is false, preschoolers continue to act on that information, for example, by following an adult's misleading indication of the location of a hidden object (Couillard and Woodward 1999; Jaswal et al. 2010). These deferential reactions lend support to the long-standing assumption that young children are credulous—disposed to trust claims made by other people even when those claims run counter to their own convictions or intuitions.

Contrary to this assumption, we argue that children are not prone to indiscriminate credulity. Instead, they engage in what we will refer to as selective trust. As just documented, young children do accept information from others, even when it runs counter to their own observations and intuitions. Nevertheless, when they meet informants who make conflicting claims they do not endorse both claims. They typically endorse those made by one informant rather than the other. In particular, they use two guiding principles or heuristics. They are inclined to accept the claims of informants with whom they have a social connection over those made by strangers. Second, they are inclined to accept the claims of informants who have proven well-informed rather than ill-informed. We describe the evidence for these two heuristics and then ask what children do when the two heuristics are placed in opposition. Whom do young children endorse if a relative stranger appears to be better informed than someone they know well?

Having reviewed the available findings, we consider their implications for children's metacognitive abilities. More specifically, we weigh up two possible interpretations. One possibility is that when children select among informants, such selectivity necessarily implies a capacity for metacognition, however limited or basic. A second possible interpretation is that children might initially select among informants, irrespective of any metacognitive capacity that they possess. On this argument, it is only when children begin to select among informants in terms of how well informed those informants are that it is legitimate to speak in terms of metacognition.

Trusting familiar informants

To find out if young children are selective when they encounter conflicting claims made by a familiar and an unfamiliar informant, we tested 3-, 4-, and 5-year-olds in two different daycare centres (Corriveau and Harris 2009a). In the presence of a familiar caregiver from their own centre and an unfamiliar caregiver from the second centre, children were presented with a series of novel objects obtained from the hardware shop. Because we tested children from both centres, approximately half regarded one caregiver as familiar and the other as unfamiliar. The children could ask for information about the name or function of the object from either caregiver. No matter which woman children asked, both women responded by proposing different names or functions for the object and children were invited to endorse one or the other. Figs 12.1 and 12.2 show the findings from the two centres. Both tell essentially the same story. All three age groups preferred to seek and accept information from the caregiver with whom they were familiar.

In a later study, we asked how far children trust the information offered by their mother as compared to a stranger (Corriveau et al. 2009a) and whether the level of trust varies with attachment status. When children were approximately 15 months old, they had been categorized as secure, ambivalent, or avoidant in their relationship to their mother based on their behaviour in the Strange Situation[1] (Ainsworth et al. 1978). At 4 years of age, children's selective trust was

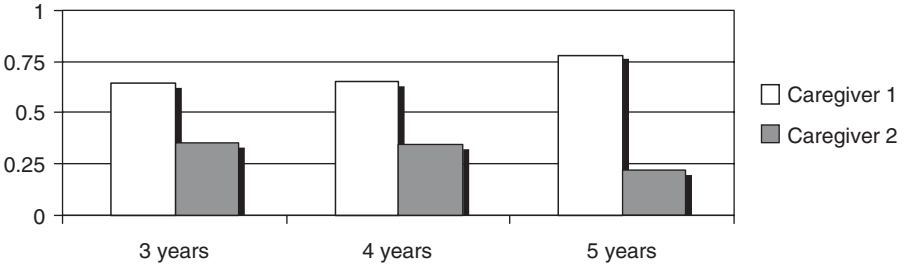


Fig. 12.1 Proportion of choices directed at each caregiver by 3-, 4-, and 5-year-olds in Centre 1.

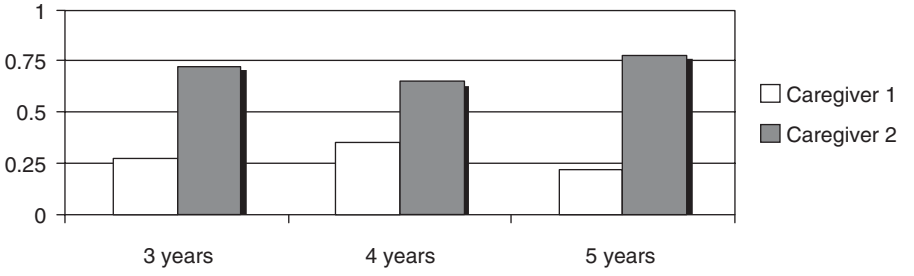


Fig. 12.2 Proportion of choices directed at each caregiver by 3-, 4-, and 5-year-olds in Centre 2.

[1] The Strange Situation consists of a series of episodes in the course of which infants are briefly separated from their mother and eventually reunited. The infants' behaviour is coded to assess how they cope with the separation and how readily they are reassured when the mother returns. A large body of findings suggests that children show a relatively stable set of reactions toward a given caregiver both during and after the separation, reactions that reflect expectations that infants build up about how reliable the caregiver is as a source of comfort and reassurance.

assessed using a procedure similar to the one just described for the two daycare centres. Children were invited to ask for and accept information about the names or functions of unfamiliar objects from either their mother or from a relatively unfamiliar stranger. Fig. 12.3 shows the proportion of choices that children in each of the three attachment groups directed at their mother as compared to the stranger. Inspection of Fig. 12.3 shows that, overall, we replicated the pattern found in the earlier study. Children preferred to trust the information supplied by a familiar informant—their mother—as compared to that supplied by a stranger. Nevertheless, Fig. 12.3 also shows that the strength of that preference varied across the three attachment groups. It was an unreliable trend among children with an avoidant relationship. On the other hand, it was a systematic preference among children with a secure or ambivalent relationship.

Apparently, children's trust in an informant, including a highly familiar informant such as their mother, is moderated by their history of interaction with her. It is too early to say what particular aspects of that interaction are critical. Still, in line with the classic tenets of attachment theory, it is plausible that mothers vary in their responsiveness and children come to notice and encode that variation and respond accordingly. For example, based on past experience, avoidant children might have come to the conclusion that their mothers are relatively unresponsive as informants. As a result, avoidant children show no particular preference for the information that she can supply as compared to that of a stranger. Secure children might be confident about their mother's responsiveness and systematic in seeking and accepting the information that she provides. Ambivalent children might be especially, indeed uncritically dependent on the information supplied by their mother, especially in comparison to that supplied by a stranger.

One other, related point is worth emphasizing. Evidently, mere familiarity with an informant is no guarantee that the information she supplies will be preferred. It is tempting to draw that conclusion from the study with children's caregivers in daycare (Figs 12.1 and 12.2) but the findings for the avoidant children (Fig. 12.3) show that such a conclusion would be mistaken. Even though avoidant children were obviously familiar with their mother, they did not prefer the information she supplied to that of a stranger. By implication, when children build up trust in a caregiver over repeated encounters, they are not just accumulating feelings of familiarity—they are also building up a social or emotional connection.

Approximately 12 months later, when the children were 5 years of age, we again tested their reactions to their mother as compared to a stranger. The children were presented with pictures of animal hybrids that were of two different types. One type consisted of symmetric hybrids: each hybrid resembled two different animals—such as a cow and a horse—to the same degree. We anticipated that when children heard their mother categorize the hybrid in one way—'That's a horse'—and the stranger categorize it in another way—'That's a cow'—they would respond as they had done with the novel, hardware objects. Even if, objectively speaking, each categorization

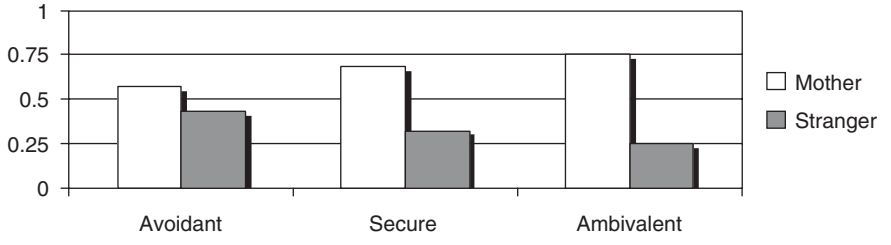


Fig. 12.3 Proportion of choices directed by 4-year-olds at their mother versus a stranger as a function of attachment classification (unfamiliar objects).

was equally consistent with the available perceptual evidence, we expected children to favour the categorization supplied by their mother. We also expected the strength of that preference to vary depending on the nature of children's attachment to their mother. Inspection of Fig. 12.4 shows that both of these expectations were borne out. Overall, children preferred to seek and accept information from their mother rather than from the stranger. Nevertheless, the strength of that preference varied with the child's attachment status. It was absent among avoidant children, systematic among secure children, and very strong among ambivalent children.

These results were encouraging. They confirmed the pattern that we had observed initially, showing that it was robust even though children were almost 1 year older, thereby further extending the time that had elapsed between children's assessment in the Strange Situation and our test of selective trust.

The second set of hybrids was asymmetric: they resembled two different animals to different degrees. For example, for 75% of its perceptual features an asymmetric hybrid might resemble a squirrel but for the remaining 25% of its features it might resemble a rabbit. The mother always named the animal in terms of the less plausible category (e.g. 'That's a rabbit') whereas the stranger named the animal in terms of the more plausible category (e.g. 'That's a squirrel'). We anticipated two different possible outcomes. First, suppose that children did not encode the balance of the perceptual evidence. For example, they might simply note the resemblance to both a rabbit and a squirrel but fail to notice that overall the evidence pointed to its being a squirrel rather than a rabbit. If that were the case, the pattern of results should be the same as displayed in Figs 12.3 and 12.4. Alternatively, suppose that children did notice the asymmetry and left to their own judgement would be more likely to categorize the hybrid as a squirrel than a rabbit. To the extent that children weigh that perceptually plausible categorization against the alternative categorization proposed by their mother, we might expect them to display less confidence in the claim made by their mother. Moreover, to the extent that all children, no matter what their attachment status, are likely to have similar perceptual intuitions about the hybrid, we might reasonably expect that reduction in confidence to be roughly comparable across all three attachment groups.

Inspection of Fig. 12.5 shows that the pattern of results fits the second proposal not the first. Children do seem to notice that they are dealing with asymmetric rather than symmetric hybrids. This perceptual intuition undermines confidence in their mother's claims and this reduction is of approximately the same magnitude for all three attachment groups: compare Figs 12.4 and 12.5. More broadly, the different pattern of findings obtained with the asymmetric as compared to the symmetric hybrids implies that children's reactions to the claims made by others depend on their own convictions about what they see. When they are uncertain—as they presumably were in the

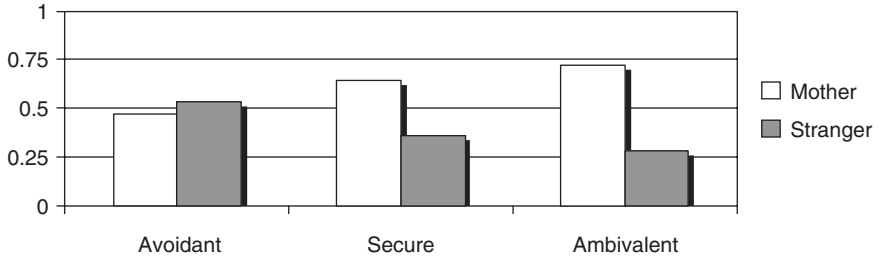


Fig. 12.4 Proportion of choices directed by 5-year-olds at the mother versus a stranger as a function of attachment classification (symmetric hybrids).

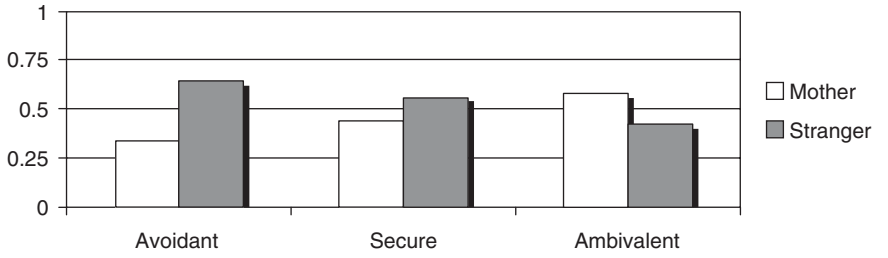


Fig. 12.5 Proportion of choices directed by 5-year-olds at the mother versus a stranger as a function of attachment classification (asymmetric hybrids).

case of the symmetric 50/50 hybrids—they readily yield to a trusted informant such as their mother. On the other hand, when provided with more counter evidence– as with the asymmetric 75/25 hybrids, they are less likely to defer to an ordinarily trusted informant.

One final result is worth putting in context. In a series of experiments, Vikram Jaswal has also assessed the extent to which children defer to an informant when categorizing objects. He reports that when presented with asymmetric hybrids (for example, the rabbit–squirrel), and told by a single adult informant that the hybrid belongs to the less likely category, children are inclined to defer to the adult. Indeed, they defer despite the fact that, when left to their own devices, they categorize the hybrid in terms of the more likely category. This deference is less evident among 4-year-olds than 3-year-olds but even 4-year-olds readily defer if the adult signals that what he or she is claiming might seem unlikely ('You're not going to believe this but …') (Jaswal 2004). Based on these findings, we might have expected that most children would defer to their mother's claims even when presented with asymmetric hybrids. Indeed, to the extent that Jaswal found that most children deferred to an unfamiliar adult who identified the hybrid as belonging to the less likely category, we might have expected children in our study to defer even more to their own mother who also identified the hybrid as belonging to the less likely category.

The children in our study were somewhat older than those tested by Jaswal (2004). As children get older, they might be increasingly sceptical of counterintuitive claims, even those made by a trusted informant such as their mother. In addition, however, a procedural change may have played an important role. In the studies conducted by Jaswal (2004), children heard the unexpected categorization proposed by a single, unfamiliar adult. By contrast, in the study that we conducted, children heard two informants propose conflicting categorizations—their mother proposed a less plausible categorization but a stranger proposed a more plausible categorization. Arguably, the children in our study did not simply weigh their own perceptual judgement against the proposal made by their mother, they were also bolstered in making their own judgement by the fact that it coincided with that of the stranger. In short, children are more confident of their own perception-based conviction if another person—even a stranger—agrees with them. Indeed, we might reasonably speculate that had their mother's proposal coincided with their perception-based conviction whereas the stranger presented children with a less plausible categorization, children would have displayed a very strong preference for the more plausible categorization.

These findings underline the claim that in assessing children's credulity, we should not simply try to find out how far they defer to the judgement of another person. We need to ask how they select among the conflicting claims of various informants. In the next section, we look at the question of whether preschoolers select among informants on epistemic as opposed to socioemotional grounds.

Trusting knowledgeable informants

To examine this dimension of children's selective trust, we have adopted the following basic paradigm. First, children are given information about the differential knowledge or reliability of two informants in a familiarization period. Then, in a test period, an unfamiliar object is presented and children are given an opportunity to seek and accept information about it from one or other of the two informants. We measure the extent to which children choose to rely on the more knowledgeable of the two informants. In one series of experiments, we assessed children's ability to distinguish between a knowledgeable and an ignorant informant. In the familiarization period, children were introduced to two informants—one named each of three common objects accurately whereas the other admitted to not knowing their names (Koenig and Harris, 2005, experiment 2). Because the objects were familiar, the children could confirm for themselves that one informant knew the right names for the objects even if the other claimed ignorance. Before and after the ensuing test period, children were asked to make an explicit judgement about the relative knowledge of the two informants. More specifically, they were asked to judge who was 'not very good at answering the questions' about the names of the objects. In the test period, children were first shown another familiar object and asked to predict what the two informants would say about it. They were then shown three novel objects whose names they did not know, invited to ask one of the informants what each novel object was called, and after each had suggested a different name for the novel object, to endorse one of the two supplied names.

Overall, both age groups proved to be remarkably good at judging, predicting and utilizing the difference between the two informants (see Fig. 12.6). Thus, in answering the explicit judgement questions, children reliably picked out the informant who was 'not very good' at answering the questions. In the prediction trials, they anticipated that one informant would name the object accurately whereas the other would acknowledge ignorance or make a mistake. When given an opportunity to ask for information, they preferred to ask the knowledgeable as opposed to the ignorant informant. Finally, when given an opportunity to endorse the name supplied by one informant or the other, they tended to endorse the name supplied by the knowledgeable informant—although this selective pattern of endorsement was weaker among 3-year-olds as compared to 4-year-olds.

Is children's selective trust confined to object naming—the domain in which the two informants had displayed differential knowledge? Alternatively, do they also display selective trust if the two informants offer information about a different domain—for example, object functions—as

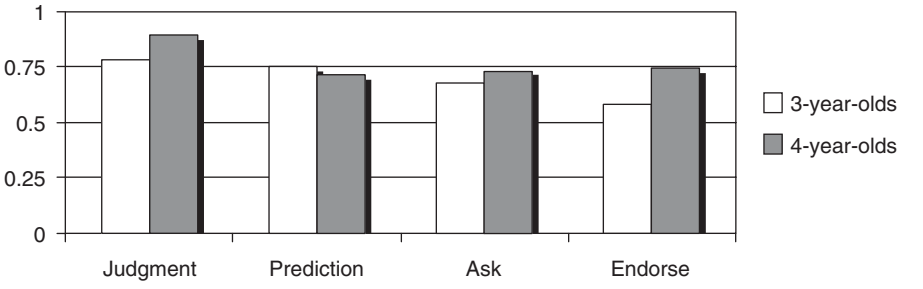


Fig. 12.6 Proportion of correct responses as a function of question type and age. (Koenig and Harris, 2005, experiment 2). Reproduced from Melissa A. Koenig and Paul L. Harris, Preschoolers Mistrust Ignorant and Inaccurate Speakers, *Child Development*, 76, 1261–77 © 2005, John Wiley and Sons, with permission.

well as object names? This issue was pursued in a follow-up experiment (Koenig and Harris 2005, experiment 3).

Children were again introduced to two informants, one who proved accurate and one who proved ignorant in naming familiar objects. They were then shown four unfamiliar objects. For two of the four objects, children were invited to seek help concerning their names. For the other two objects, they were invited to seek help concerning their functions. Fig. 12.7 displays the findings for 3- and 4-year-olds. In the explicit judgement trials, children in both age groups were again very accurate in picking out the person who was 'not very good' at answering the questions. Moreover, as before, when the informants proposed conflicting names children typically endorsed the name offered by the knowledgeable informant. Children also displayed a very similar pattern with respect to object functions—they preferred to ask for help from and endorse the function modelled by the more knowledgeable informant.

By implication, having learned about the accuracy with which the knowledgeable informant could name objects, children did not make a very narrow assessment of her knowledge. They also took her to be knowledgeable about object functions. In due course, we will revisit the question of how broad or narrow children's attributions are.

Selective trust might be quite easy for young children to display when they are confronted by an ignorant as compared to a knowledgeable informant—especially when one informant explicitly admits ignorance. How do they react when the two informants vary in a less explicit fashion? To explore this issue, we introduced children to one informant who was accurate and another who was inaccurate in stating the name or the properties of familiar objects. In the subsequent test phase, the two informants supplied information about the names or properties of unfamiliar objects. In these initial experiments, the typical pattern was for 4-year-olds to display selective trust by asking for and endorsing information from the accurate informant whereas 3-year-olds were less systematic (Koenig et al. 2004; Koenig and Harris 2005).

Subsequently, we made various procedural changes designed to facilitate children's recognition and retention of the fact that one informant was accurate whereas the other was inaccurate. The number of familiarization trials was increased from three to four—with the accurate informant naming all four objects correctly and the inaccurate informant naming all four objects incorrectly. The two informants were made more distinctive from one another in terms of clothing. Finally, they remained seated in the same place throughout the familiarization and the test period to facilitate children's ability to re-identify each informant from one phase of the experiment to the next (Pasquini et al. 2007).

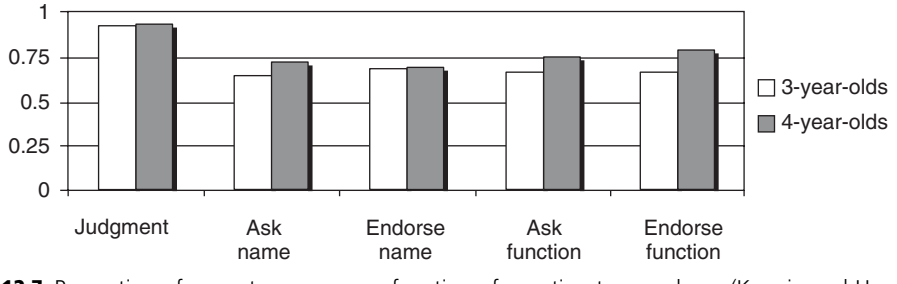


Fig. 12.7 Proportion of correct responses as function of question type and age (Koenig and Harris 2005, experiment 3). Reproduced from Melissa A. Koenig and Paul L. Harris, Preschoolers Mistrust Ignorant and Inaccurate Speakers, *Child Development*, 76, 1261–77 © 2005, John Wiley and Sons, with permission.

These modifications proved helpful. As Fig. 12.8 shows, although 3-year-olds continued to perform less accurately than 4-year-olds, both age-groups performed above chance on three types of probe: they explicitly judged the inaccurate informant to be 'not very good' at answering questions during the familiarization period; they asked for help from the accurate as opposed to the inaccurate informant; and they endorsed the information that she supplied.

Children were also assessed for their ability to solve a standard false belief task involving a misleading container. Overall, they performed quite poorly: 3-year-olds performed below chance whereas 4-year-olds were more mixed with some performing correctly and others not, so that group performance was at chance. However, as just noted, this did not prevent either group from displaying selective trust in the more accurate informant. A clear implication of this conjunction of findings is that correct performance on a standard false belief task is *not* a prerequisite for selective trust in a more accurate informant.

Summing up the findings so far, preschoolers are quite sensitive to variation between informants in their knowledge. If one informant is consistently knowledgeable or accurate whereas the other is either consistently ignorant or inaccurate, they display selective trust. They appropriately judge the reliable informant to be better at answering questions; they anticipate how each informant will describe an unfamiliar object; they seek information from the more reliable informant; and they selectively endorse the information that they receive from that informant.

In the experiments described so far, each of the two informants behaved in a consistent fashion. One was consistently reliable whereas the other was consistently unreliable. Outside of the laboratory, however, informants are rarely so consistent. They are likely to display a mix of accuracy and inaccuracy, or truth and error. Despite this mix, we nonetheless judge some informants to be generally reliable whereas we are dubious about others. By implication, we form a global impression of someone's trustworthiness—weighing their overall accuracy against their occasional inaccuracy. Do preschoolers display a similar tendency? More specifically, when faced with informants who are less than fully consistent, do they form a global impression of their trustworthiness? To examine this issue, we included two further conditions in the experiment just described. Recall that in one condition children were introduced to one informant who was accurate across all four trials and one informant who was inaccurate across all four trials. We may refer to this as the '100% vs. 0%' condition. In two further conditions, the accurate and/or the inaccurate informant were not fully consistent. In one condition ('75% vs. 0%') one informant was accurate on three of the four trials and the other was consistently inaccurate. In a second condition ('100% vs. 25%'), one informant was consistently accurate whereas the other was inaccurate on three of the four trials.

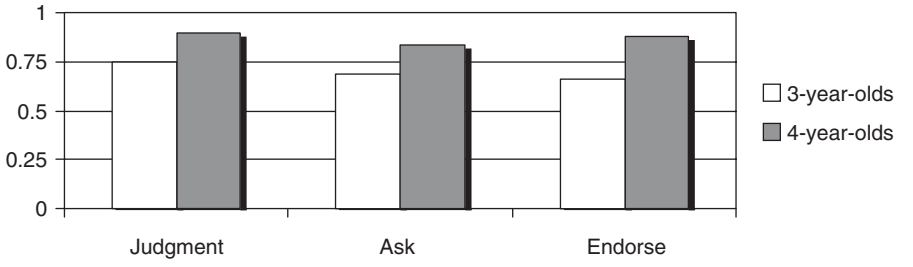


Fig. 12.8 Proportion of correct responses by age and type of question (Pasquini et al. 2007, experiment 1; 100% vs. 0% condition). Reproduced from Pasquini, E. S., Corriveau, K., Koenig, M., and Harris, P. L., Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, 43, 1216–26 © 2007, The American Psychological Association.

In the two new conditions, 4-year-olds were systematic across all three types of test question. They picked out the person who was 'not very good' at answering the questions; they sought help from the other informant; and when given suggestions by both informants, they typically endorsed the more accurate informant. These results show that 4-year-olds 'forgive' occasional errors. Even though the more accurate informant was not consistently accurate in the 75% vs. 0% condition, 4-year-olds appeared to overlook that error and treat her as the more trustworthy informant. Conversely, even though the less accurate informant was not consistently inaccurate in the 100% vs. 25% condition, 4-year-olds still treated her as the less trustworthy informant. Stated simply, 4-year-olds appear to recognize and accept that informants will display occasional inconsistency—they will sometimes be accurate and sometimes inaccurate—but they prefer those who, on balance, are more rather than less accurate.

The results for the 3-year-olds were less simple but provocative. First, in the 100% vs. 25% condition, although they were somewhat less accurate than 4-year-olds, they too were systematic in their answers to all three types of questions. On the other hand, in the 75% vs. 0% condition, they behaved in an essentially random fashion across all three test questions. Note that the more accurate informant in this condition made only a single error. By implication, 3-year-olds are 'unforgiving'. They treat an informant making a single error as no more trustworthy than someone making multiple errors.

Further evidence for the different stance of 3- and 4-year-olds emerged in a follow-up experiment. We compared children's performance in two conditions: 75% vs. 0% and 75% vs. 25% (Pasquini et al. 2007, experiment 2). If 4-year-olds can monitor the overall balance of accuracy versus inaccuracy, they should display selective trust in both conditions but if 3-year-olds are unforgiving of single errors, they should fail to display selective trust in either. The results fit these expectations. Overall, 4-year-olds displayed selectivity in both conditions but 3-year-olds did so in neither. Fig. 12.9 shows the results (collapsed across judgement, ask and endorse trials) for the three conditions of the initial study (100% vs. 0%; 100% vs. 25%; 75% vs. 0%) and the two conditions of the follow-up study (75% vs. 0%; 75% vs. 25%). Inspection of Fig. 12.9 confirms that 4-year-olds performed above chance in all five conditions whereas 3-year-olds performed above chance in only two conditions—those in which one informant was 100% accurate.

In all the experiments described so far, one informant proved to be relatively well-informed and the other ill-informed. In principle, therefore, children might not have reduced trust in the ill-informed speaker. They might have increased trust in the well-informed speaker. To examine

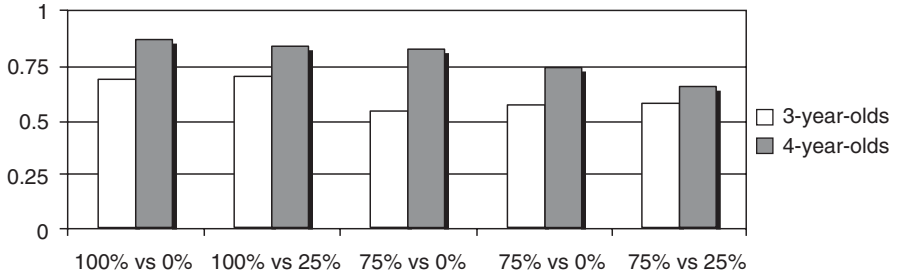


Fig. 12.9 Proportion of correct responses by age and condition (Pasquini et al. 2007, experiments 1 and 2). Reproduced from Pasquini, E. S., Corriveau, K., Koenig, M., and Harris, P. L., Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, 43, 1216–26 © 2007, The American Psychological Association.

this possibility, Corriveau et al. (2009b) tested 3- and 4-year-olds in three conditions. The accurate–inaccurate condition was similar to the 100% vs. 0% condition just described. In the familiarization period, one of the informants consistently named familiar objects accurately whereas the other informant consistently named them inaccurately. In the accurate–neutral condition, one of the informants consistently named familiar objects accurately whereas the other made only neutral or non-committal remarks—'Oh, look at that'. Finally, in the inaccurate–neutral condition, one of the informants consistently named familiar objects inaccurately whereas the other made only neutral or non-committal remarks—'Oh, look at that'.

Both 3- and 4-year-olds preferred the accurate informant in the accurate–inaccurate condition, consistent with earlier findings. In addition, both age groups preferred the neutral informant in the inaccurate–neutral condition suggesting that when one of two speakers makes mistakes that is sufficient to elicit mistrust. A clear age change emerged in accurate–neutral condition. Four-year-olds preferred the accurate informant whereas 3-year-olds were not selective.

These data suggest that 4-year-olds keep a fairly precise and comprehensive record of their informants, building up trust in those who have proven accurate and mistrust in those who have proven inaccurate. Three-year-olds, by contrast, appear to focus in a narrower fashion on inaccuracy. If an informant makes a mistake—even a single mistake—they become mistrustful. If both informants make mistakes even with differential frequency, or if one informant is accurate and the other non-committal, they invest no more trust in the one than the other. By implication, 3-year-olds are solely on the look out for mistakes. Whether confronted by a single mistake or by several, their reservoir of trust in that person is depleted.

A plausible underpinning for this particular developmental change is the improvement in children's understanding of false belief that is widely observed between 3 and 5 years (Wellman et al. 2001). From that perspective, younger children think of the mind as a passive recorder or copier of events (Chandler 1988; Taylor et al. 1991). So, for 3-year-olds a source is trustworthy when he or she has been 'in contact' with the relevant information. By contrast, older children possess an interpretative theory of mind in which representations may be detached from, or even inconsistent with, their referent, so that the source may be more or less correct. According to this interpretation, children who grasp the potential for false beliefs—typically children aged 4 years and upward—not only withdraw credit in the case of false statements they also tender credit in the case of true statements. By contrast, 3-year-olds typically fail to grasp the potential for false beliefs. Hence, although they withdraw credit in the case of false statements they take true statements for granted. So, extending the argument made earlier: an understanding of false beliefs is not a precondition for mistrusting an inaccurate speaker. As noted earlier, children who fail standard false belief tasks, including 3-year-olds, are able to do that. On the other hand, an understanding of false beliefs may well be a precondition for the augmentation of trust in an accurate speaker.

Weighing socioemotional and epistemic signs of trustworthiness

So far, we have identified two quite different strategies that 3- and 4-year-olds use to select among their informants. First, they use a relational strategy. They prefer to gather and receive information from an informant with whom they have an established relationship—at least, provided it is not avoidant. Second, they use a more epistemic strategy. They prefer to gather and receive information from someone who has proven reliable. Thus, they are mistrustful of informants who have indicated their unreliability, either by acknowledging their ignorance or by making obvious and easily identifiable mistakes. Indeed, 4-year-olds seem especially attuned to differences in

accuracy because when someone proves accurate they do not take that accuracy for granted but strengthen their trust in that person.

What happens when these two strategies, the relational and the epistemic, are placed in conflict with one another? For example, how do preschoolers respond when they encounter a familiar informant who makes mistakes? They might ignore the mistakes and continue to invest selective trust in the familiar informant. Alternatively, they might attend to the mistakes and come to mistrust the informant despite his or her familiarity. Still, a third possibility is that there is an age change in the preschool period with younger children attending more to familiarity and older children to accuracy. To assess these three possibilities, Corriveau and Harris (2009a) extended the experiment described earlier involving two preschool caregivers, one familiar and one unfamiliar. The complete experiment had three phases: pretest trials, accuracy trials and post-test trials. As described earlier, 3-, 4-, and 5-year-olds were shown unfamiliar objects in the pretest trials and given the opportunity to learn about them from the caregivers. Recall that children in all three age groups and in each childcare centre displayed a preference for the more familiar caregiver (Figs 12.1 and 12.2).

In the subsequent accuracy trials, children received information about the accuracy of the two caregivers. They were shown a set of familiar objects whose names they knew. Half the children heard the familiar caregiver name these objects accurately and the unfamiliar caregiver name them inaccurately. The remaining children heard the reverse arrangement: the familiar caregiver named them inaccurately whereas the unfamiliar caregiver named them accurately.

In post-test trials, children were shown four unfamiliar objects and were given Ask and Endorse probes akin to those in the pretest trials. Thus, we could check whether children's initial preference for the familiar caregiver was either strengthened or undermined depending on her behaviour in the accuracy trials. Fig. 12.10 shows the proportion of times that children continued to select the familiar informant after receiving information about her relative accuracy during the accuracy trials.

In the post-test trials, 5-year-olds were very sensitive to the information provided during the preceding accuracy trials. If the familiar informant had been accurate, they displayed a marked preference for her but if she had proven inaccurate, they switched, and displayed a preference for the hitherto unfamiliar—but accurate—informant. Four-year-olds also proved sensitive to the information provided during the accuracy trials. Like the 5-year-olds, they displayed a marked preference for the familiar informant if she had proven accurate but no systematic preference for

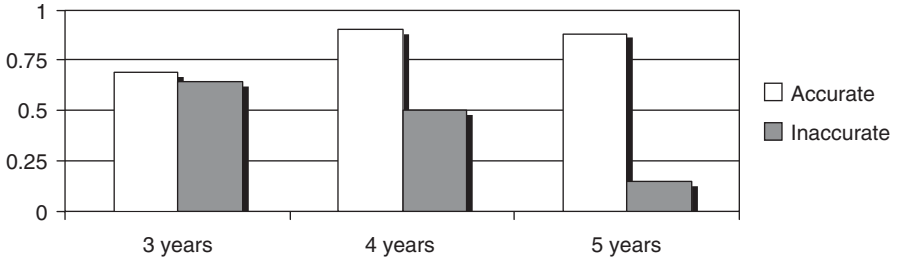


Fig. 12.10 Proportion of times children selected the more familiar informant (collapsing across ask and endorse probes) by age and behaviour of the familiar informant (accurate versus inaccurate) during the accuracy trials. Reproduced from Kathleen Corriveau and Paul L. Harris, Choosing your informant: weighing familiarity and recent accuracy, *Developmental Science*, 12, 426–37 © 2009, John Wiley and Sons, with permission.

her if she had proven inaccurate. Finally, the 3-year-olds were the least affected by accuracy information. Ignoring the feedback they had received in the accuracy trials, they maintained the preference for the familiar informant that they had displayed during the pretest trials.

Clearly, this pattern of findings supports the third possibility: younger preschoolers attend more to familiarity and older preschoolers attend more to accuracy. But how can we explain the age change? Three possible explanations can be quickly ruled out. First, it might be argued that 3-year-olds did not notice the mistakes made by the unfamiliar informant during the accuracy trials. This is implausible because, as we saw earlier, in comparable experiments in which an unfamiliar informant made mistakes, 3-year-olds ended up mistrusting the person who made them (Pasquini et al. 2007; Corriveau and Harris 2009b).

A variant on this explanation is similarly problematic. It could be argued that 4-year-olds, and certainly 5-year-olds, typically have an understanding of false belief. As argued earlier, they might be more appreciative than 3-year-olds of the fact that an informant has proven accurate. Hence, 4- and 5-year-olds might be more swayed by the information provided during accuracy trials, noting not just who is inaccurate but also who is accurate. Again, however, this argument overlooks the fact that 3-year-olds do differentiate between an accurate and an accurate informant when both are unfamiliar (Pasquini et al. 2007; Corriveau and Harris, 2009b).

Another possibility is that 3-year-olds weigh the few mistakes they have witnessed on the part of the familiar informant against a much longer history of interaction in which she has, presumably, proven accurate and they therefore discount those few errors. This line of explanation allows for the fact that 3-year-olds do differentiate between an accurate and an inaccurate informant when both are unfamiliar, and indeed are relatively unforgiving of a single error, but in the case of an unfamiliar informant, no prior history of interaction is available to serve as a counterweight to recent evidence of inaccuracy. However a similar argument would seem to apply just as forcefully to 4- and 5-year-olds. After all, they are likely to have had at least as long a history of interaction with the familiar informant as 3-year-olds. Yet despite their potentially deeper reservoir of accumulated trust, 4- and 5-year-olds did alter their pattern of trust in the wake of inaccuracy on the part of the familiar informant.

A more plausible explanation is that the findings reflect an important developmental shift in the relative weight that children attach to two different indices of trustworthiness: familiarity, or more broadly a feeling of social connection on the one hand and epistemic competence on the other. The data suggest that even though 3-year-olds can assess epistemic competence—they prefer accurate and knowledgeable informants to those who are inaccurate and ignorant—when asked to weigh that index of trustworthiness against familiarity, they attach more weight to social connection. The reverse is true for 5-year-olds. Other things being equal, they prefer to put their questions to, and accept information from, a familiar informant. However, when asked to weigh familiarity against epistemic competence—as indexed, for example, by accuracy—they prefer an unfamiliar but accurate informant to one who is familiar but inaccurate.

One concern about this line of explanation is that it could be regarded as a simple re-description of the findings. However, it is important to emphasize that, as formulated, the explanation goes beyond an account of the pattern of results depicted in Fig. 12.10. The implication is that a variety of cues that might promote a sense of social connection will be increasingly trumped by epistemic competence in the course of the preschool years. For example, recent evidence shows that preschoolers prefer to learn from an informant who speaks with a native accent versus a foreign accent (Kinzler et al. 2010). We assume that this preference is driven by feelings of social connectedness toward someone perceived as a member of the ingroup. As such, we anticipate that this preference would be relatively persistent among 3-year-olds. Thus, even if an informant with a native accent proved inaccurate, they would favour him or her over an informant with a

foreign accent. By contrast, among 4- and 5-year-olds epistemic competence would be weighed more heavily. Recent findings provide support for exactly this age change (Corriveau et al. submitted).

Implications for metacognition

We have argued that young children use two different heuristics for selecting among informants. First, they prefer to learn from those who elicit feelings of social connectedness. Second, they prefer to learn from someone who displays epistemic competence by making well-informed claims. When these two heuristics are placed in conflict with one another, younger children—3-year-olds—favour social connectedness whereas older children—5-year-olds—favour epistemic competence.

We may now step back and consider the relationship between children's selective trust and the development of metacognition. We will consider whether children select among their informants by relying on either analytic metacognitive judgements or on metacognitive feelings (Koriat 2000; Proust 2007). At first sight, the developmental shift that we have described strongly suggests that children increasingly do make metacognitive evaluations in the sense that they appraise the epistemic standing of their informants. Still, this conclusion is open to doubt. Arguably, even though 4- and 5-year-olds choose among their informants in terms of relative accuracy it could be argued that they do so without any systematic recourse to metacognitive reflection. We will consider arguments for and against this conservative conclusion.

To the extent that children select among their informants in terms of social connectedness, one can plausibly argue that it is unnecessary to invoke any role for metacognition in that selection. On this sceptical argument, children give no thought to the possibility that those with whom they have a social connection offer more reliable or trustworthy information. Instead, they have an early and non-reflective bias to encode and retain information from their nearest and dearest. In much the same way, recent evidence suggests that non-human primates are biased to emulate models with greater prestige. Here too, it is plausible that such a bias is not guided by any consideration of the relative reliability of the information that high-ranking individuals provide as compared to low-ranking individuals. Admittedly, this is not to explain the origins of such a bias. Arguably, it is an innate bias that is built into social learning whether it is undertaken by children or by non-human primates. Alternatively, it is a by-product of information-processing biases that are likely to ensue from social preferences. For example, it is feasible that information delivered by a familiar attachment figure or by a high-ranking model is processed more extensively or deeply because such a source typically receives preferential attention. Still, even pending a fuller explanation of the basis for children's preference for information supplied by a familiar informant, it is unlikely that we need to infer any metacognitive basis for that selectivity.

However, it is worth discussing a possible caveat to this sceptical conclusion. Recall that when their mother and a stranger provided conflicting information, children's reactions varied in two ways. First, children's trust in their mother as compared to the stranger varied depending on the type of attachment that they had to her. But second, and more relevant to a potential role for metacognition, no matter what their attachment history, children were less prone to trust the claim provided by their mother when the hybrid creatures were asymmetric, i.e. in those cases when the categorization proposed by the mother was less consistent with the available perceptual evidence than the categorization proposed by the stranger. Recall that this was evident in the pattern of trust invested in the mother in Figs 12.3 and 12.4 as compared with the pattern in Fig. 12.5.

An initially plausible interpretation of this variation is that children engage in a metacognitive assessment of their level of confidence in the categorization that they themselves believe to

be likely. Indeed, animal studies have demonstrated that such uncertainty evaluations exist even in monkeys and dolphins (Smith 2009). Moreover, recent developmental studies have suggested that 3-year-olds also have an implicit access to their knowledge states (Balcomb and Gerken 2008). Thus, children may have looked at the asymmetric squirrel–rabbit hybrid and judged themselves to be more confident of its being a squirrel than a rabbit. Hence, when their mother called it a rabbit but the stranger called it a squirrel, their differential confidence in those two categories inclined them to accept the stranger's proposal—at least more so than in cases when the mother's claim had been as consistent with the available perceptual evidence as the stranger's. However, it is also possible that children make no such metacognitive assessment of their feelings of confidence in a particular categorization or even if they are capable of such an assessment, make no use of it in weighing up the conflicting proposals. Instead, it is possible that two alternative categorizations—'squirrel' and 'rabbit' each come to mind but the strength or availability of those two representations differs. As a result, one representation is more susceptible to endorsement and retention than the other. Thus, when the mother and the stranger make conflicting proposals, the proposal that coincides with the stronger representation has a greater chance of survival, and of being adopted and endorsed by the child. On this admittedly cautious view, even if children were capable of stating their relative confidence in the two different categorizations ('I'm sure it's a rabbit – I doubt it's a squirrel'), it could still be the underlying strength or availability of the two representations that actually determines whether the mother or the stranger's claim is accepted.

In sum, reviewing the evidence that was mustered in the first section, there is, for the time being anyway, no compelling reason to conclude that children's selective learning from particular informants implies that they engage in any metacognitive assessment of the relative trustworthiness of different informants or the relative plausibility of their own intuitions. Children do undoubtedly display selective trust but, for the time being, there is no firm evidence that it is guided by any metacognitive reflection on either the knowledge of their informants or their own knowledge.

We may now consider the evidence discussed in the second section. To the extent that children appraise informants in terms of the accuracy with which they have named familiar objects, does this imply some type of metacognitive reflection? At first sight, the evidence would seem to call for a positive answer. Recall that both 3- and 4-year-olds consistently preferred to learn from an informant who had proven more accurate in naming familiar objects. A plausible interpretation of this selectivity is that children judge that the hitherto more accurate informant is more knowledgeable. To that extent, such selective trust would imply a capacity for metacognition in the sense that children make judgements about the differential knowledge base of the two informants and accept information from the more knowledgeable informant—even if they give little thought to the mental processes by which an informant retrieves information from his or her knowledge base in answering a given question.

However, there are again reasons for caution. First, when the less accurate informant consistently misnames familiar objects, children may conclude, particularly when there is no obvious reason for the informant's errors, that the less accurate informant is simply deviant. Lucas and Lewis (2010, p. 168) formulate this caution as follows: 'It may be that the expectation for correct labelling is so ingrained in young children that a violator is perhaps more likely to be viewed as globally incompetent or bizarre, rather than misinformed'. Hence, on this interpretation, children do not make a genuinely metacognitive appraisal of the less accurate informant. Instead, they make a more generic appraisal (she is 'globally incompetent'). Alternatively, they focus on her repeated deviation from a social norm (she is 'bizarre'). Both of these possibilities warrant consideration because preschool children are indeed prone to global attributions and they are also quite sensitive to deviations from social norms (Rakoczy et al. 2008). Still, other evidence

shows that neither of these two possibilities offers a satisfactory account of all the relevant findings on selective trust.

Fusaro et al. (2011) presented preschoolers with two puppets, one who named familiar objects accurately and one who named them inaccurately. Children were then asked to make various predictions about each puppet. They predicted that the accurate puppet would be better at labelling objects but they anticipated no differences in other behavioural domains. Thus, they did not expect the two puppets to differ in terms of lifting objects, knowing what food particular animals eat, throwing basketballs into a hoop, or sharing cookies. Had children inferred that the inaccurate puppet was 'globally incompetent' they would presumably have expected him to do worse than the accurate puppet on the lifting, knowing and throwing tasks. Additionally, had children inferred that the inaccurate puppet was 'bizarre'—prone to deviate from social norms—they would presumably have expected him to share less than the accurate puppet. Indeed, in a control condition in which children were presented with two puppets who consistently differed in strength as indexed by their ability to lift four different containers, children did make global inferences: they predicted that the weaker puppet would be better not just at lifting but at labelling, knowing animal foods, throwing and sharing. In summary, these findings confirm the point that preschooler do indeed sometimes make global attributions of incompetence or non-conformity. Nonetheless, having observed two informants differ in accuracy, they make relatively narrow attributions.

Lucas and Lewis (2010) advocate the use of two criteria for demonstrating that children's selective trust involves an assessment of the knowledge states of potential informants. First, they propose that children should be provided with reasons that would explain the differential accuracy of two informants. They point out for example, that in the film *The Little Mermaid* children are introduced to a character Ariel who lives under the sea and misnames the human artefacts that happen to come her way from shipwrecks. In such a case, children would be likely to view her—appropriately enough—as lacking in a particular domain of knowledge rather than incompetent or socially deviant and to have a ready explanation for her ignorance, namely her non-human umwelt.

However, this criterion is overly stringent. Selective trust is likely to be especially useful if it is based on a metacognitive appraisal of informants' accuracy that is fast and frugal rather than probative. As adults, we readily make metacognitive inferences about people who differ in accuracy—namely that their differences in accuracy are due to differences in knowledge – even when we lack an explanation for the origin of those differences in knowledge. Similarly, when children encounter two informants who vary in accuracy, it is plausible that they attribute that variation to differences in knowledge, even when they are at a loss to explain how those differences in knowledge came about. Indeed, if children postponed mistrust in an inaccurate informant until they had an adequate explanation for the informant's inaccuracy, they might be vulnerable to all sorts of misinformation.

Consistent with this line of argument, when Koenig and Harris (2005) asked 3- and 4-year-olds to explain why one of the two informants had been inaccurate (i.e. was 'not good at answering questions') although over one-third were unable to supply an explanation, among those who did volunteer an explanation, the most frequently cited reason (12 out of 25 children) was speaker ignorance ('She didn't know the things' 'She doesn't know what they are'). Thus, even when children were given no background or life-history information that could explain the speaker's inaccuracy, ignorance was still the explanation that they favoured.

The second criterion proposed by Lucas and Lewis (2010) is a capacity for withholding trust in a selective fashion. More specifically, they propose that children be tested for their willingness to be 'forgiving'—to withhold negative assessments of an inaccurate informant in domains outside of the observed inaccuracy. Effectively, this means that children should be tested to check that

they make relatively, narrow, domain-specific attributions of lack of knowledge rather than global attributions of wide-ranging incompetence. As noted earlier, the findings of Fusaro et al. (2011) indicate that preschoolers do indeed make such narrow attributions.

One final important point has been emphasized by Einav and Robinson (2011). They underline the fact that accurate informants are not necessarily knowledgeable. They may be accurate only because they have just consulted someone else. Thus, when children infer that someone is knowledgeable on the basis of their accuracy, it would be appropriate for children to suppress that inference if there is evidence showing that the person's accuracy derives from a source other than that person's own knowledge. To examine this possibility, 4-year-olds were introduced to two puppets. One puppet named animals accurately without any help but the other named them accurately after receiving help from a third party. Subsequently, children were shown pictures of two unfamiliar animals and the puppets made conflicting claims about which animal was 'a tark'. Asked which puppet was right, children appropriately favoured the puppet whose prior accuracy was unaided.

Thus, the available evidence suggests that when children encounter two informants who differ in accuracy, they are prone to make a metacognitive inference—to conclude that the variation in accuracy reflects variation between the informants in their knowledge. Three pieces of evidence lend support to that conclusion. First, having observed that two potential informants differ in their accuracy, preschoolers expect local differences in knowledge rather than global differences in competence or social conformity. Second, when children are invited to explain those differences in accuracy they frequently and explicitly attribute them to differences in knowledge. Finally, their trust in a more accurate person is withdrawn if that person's accuracy appears to be based on help from a third party rather than their own knowledge base.

If we keep in mind the fact that 3-year-olds understand that knowing involves a certain causal relationship with a piece of information, and can monitor their own level of uncertainty (Sodian and Thoermer 2006), we can better understand why, in the previously described situations, even younger children are not prone to indiscriminate credulity. Situations in which the source has been causally linked to the relevant information and children have no contradictory perceptual information are likely to induce their trust in claims made by the source. By contrast, indications that the causal link between the source and the information has been broken will tend to discredit claims from that source.

Do these findings throw any light, however indirect, on children's understanding of the impact of informants on their own knowledge? For the time being, an agnostic answer is probably appropriate. It is certainly plausible that children seek knowledge in a selective fashion because they are aware of their own ignorance and recognize that one of the two informants can help reduce that ignorance. On the other hand, in virtually all of the experiments described in this chapter, children were prompted to ask one of the two informants and they were then invited to endorse one or the other of the claims that the two informants made. In future research, it will be important to study the conditions under which children seek knowledge from particular informants in a spontaneous fashion. We know from naturalistic studies of children's speech that they spontaneously ask many questions during the preschool period, especially when talking to a familiar caregiver (Chouinard 2007; Harris 2012). What we do not yet know is how far children understand that such information-seeking will reduce their ignorance, particularly if they put their questions to a knowledgeable informant.

Conclusions

Assuming that children do increasingly assess their informants not in terms of a social or emotional connection but in terms of a metacognitive appraisal of how knowledgeable they are, how

does such a shift come about? We may speculate about three different possibilities. First, in the course of development, children might draw up an increasingly detailed map of the way that knowledge is distributed. So long as they remain within their family circle, the limits to the knowledge of familiar adult caregivers may not be obvious. However, children may increasingly realize that their familiar caregivers are not fully conversant with every aspect of the wider world. An informant who is a relative stranger may turn out to be a more knowledgeable source of information. More generally, as young children's social horizon expands, they are likely to observe that knowledge and skill are not universal—particular informants know about particular contexts (Keil et al. 2008). On this hypothesis, children will be more or less slow to privilege epistemic competence over social connectedness depending on the breadth of their social experience.

A second possibility is that the shift is intimately connected to conceptual changes in children's understanding of knowledge and belief during the preschool years. We have emphasized that children who fail classic measures of false belief understanding will still come to mistrust an inaccurate informant. Nevertheless, it is feasible that progress in understanding the risk of false beliefs prompts children to recognize that accuracy is not automatic and should not be taken for granted. Hence, those who are consistently or predominantly accurate should be regarded as trustworthy.

Finally, the shift may have a strong maturational component. In most human societies, children's social circle gradually widens beyond the family in the course of the preschool years. Arguably, nature has built into children's cultural learning an endogenous shift in the weights that they attach to various signals of trustworthiness. Thus, more or less independent of the breadth of their social horizon or indeed of their level of conceptual development, children may become increasingly prone to compare familiar caregivers to other less familiar informants in terms of their accuracy and more broadly in terms of their epistemic competence.

References

Ainsworth, M. D. S., Blehar, M. C., Waters, E., and Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Hillsdale, NJ: Erlbaum.

Balcomb, F. K. and Gerken, L. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Science*, 11, 750–60.

Bascandziev, I. and Harris, P. L. (2010). The role of testimony in young children's solution of a gravity-driven invisible displacement task. *Cognitive Development*, 25, 233–46.

Chandler, M. (1988). Doubt and developing theories of mind. In J. W. Astington, P. L. Harris, and D. R. Olson (Eds.) *Developing theories of mind*, pp. 387–413. Cambridge: Cambridge University Press.

Chouinard, M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, 72(1), 1–121.

Clément, F., Koenig, M., and Harris, P. L. (2004). The ontogenesis of trust in testimony. *Mind and Language*, 19, 360–79.

Corriveau, K. H. and Harris, P. L. (2009a). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, 12, 426–37.

Corriveau, K. H. and Harris, P. L. (2009b). Preschoolers continue to trust a more accurate informant 1 week after exposure to accuracy information. *Developmental Science*, 12, 188–93.

Corriveau, K. H., Harris, P. L., Meins, E., et al. (2009a). Young children's trust in their mother's claims: Longitudinal links with attachment security in infancy. *Child Development*, 80, 750–61.

Corriveau, K. H., Meints, K., and Harris, P. L. (2009b). Early tracking of informant accuracy and inaccuracy. *British Journal of Developmental Psychology*, 27, 331–42.

Corriveau, K. H., Kinzler, K. D., and Harris, P. L. (submitted). Accuracy trumps accent when children learn words.

Couillard, N. L. and Woodward, A. L. (1999). Children's comprehension of deceptive points. *British Journal of Developmental Psychology*, 17, 515–21.

Einav, S. and Robinson, E. J. (2011). When being right is not enough: Four-year-olds distinguish knowledgeable from merely accurate informants. *Psychological Science*, 22, 1250–3.

Fusaro, M., Corriveau, K. H., and Harris, P. L. (2011). The good, the strong, and the accurate: Preschoolers' evaluations of informant attributes. *Journal of Experimental Child Psychology*, 110(4), 561–74.

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Cambridge, MA: The Belknap Press/Harvard University Press.

Horner, V. and Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). *Animal Cognition*, 8, 164–81.

Jaswal, V. K. (2004). Don't believe everything you hear: Preschoolers' sensitivity to speaker intent in category induction. *Child Development*, 75, 1871–85.

Jaswal, V. K. (2010). Believing what you're told: Young children's trust in unexpected testimony about the physical world. *Cognitive Psychology*, 61, 248–72.

Jaswal, V. K., Croft, A. C., Setia, A. R., and Cole, C. A. (2010) Young children have a specific, highly robust bias to trust testimony. *Psychological Science*, 21, 1541–7.

Keil, F. C., Stein, C., Webb, L., Billings, V. D., and Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive Science*, 32, 259–300.

Kinzler, K. D., Corriveau, K. H., and Harris, P. L. (2010). Children's selective trust in native-accented speakers. *Developmental Science*, 14, 106–11.

Koenig, M. and Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, 76, 1261–77.

Koenig, M., Clément, F., and Harris, P. L. (2004). Trust in Testimony: Children's use of true and false statements. *Psychological Science*, 10, 694–8.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149–71.

Lucas, A. J. and Lewis, C. (2010). Should we trust experiments on trust? *Human Development*, 53, 167–72.

Lyons, D. E., Damrosch, D. H., Lin, J. K., Simeone, D. M., and Keil, F. C. (2011). The scope and limits of overimitation in the transmission of artifact culture. *Philosophical Transactions of the Royal Society B*, 336, 1158–67.

Lyons, D. E., Young, A. G., and Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 19751–6.

Nielsen, M. and Tomaselli, K. (2010). Overimitation in Kalahari Bushman children and the origins of human cultural cognition. *Psychological Science*, 21, 729–36.

Pasquini, E. S., Corriveau, K., Koenig, M., and Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, 43, 1216–26.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 159, 271–95.

Rakoczy, H., Warneken, F., and Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*, 44, 875–81.

Smith, J. D. (2009). The study of animal metacognition. *Trends in Cognitive Science*, 13, 389–96.

Sodian, B., Thoermer, C., and Dietrich, N. (2006). Two- to four-year-old children's differentiation of knowing and guessing in non-verbal task. *European Journal of Developmental Psychology*, 3, 222–37.

Taylor, M., Cartwright, B. S., and Bowden, T. (1991). Perspective-taking and theory of mind : Do children predict interpretive diversity as a function of differences in observers' knowledge. *Child Development*, 62, 1334–51.

Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development*, 72, 655–84.

Section III

Functions of metacognition

This page intentionally left blank

Chapter 13

The subjective confidence in one's knowledge and judgements: some metatheoretical considerations

Asher Koriat

The subjective certainty in one's own knowledge

Examination of the history of early Greek philosophy reveals a shift from preoccupation with ontological questions to preoccupation with epistemological questions (Burnet 1930). Pre-Socratic Greek philosophy began by asking ontological questions—questions about the nature of the universe: what does it consist of? What is its origin? Is it infinite? When disagreements broke out, attention shifted inwards, to epistemological questions about the nature of knowledge itself: what is knowledge? How do we know? How can we be certain about our own knowledge? These questions are at the heart of present-day epistemology as well as cognitive psychology. The tension between ontological and epistemological perspectives—between asking questions about what is out there and asking questions about how we know what is out there—is not only the province of philosophy or psychology; it is today the province of modern physics as well. Some of the discussions in modern physics raise the question of whether the processes by which we acquire knowledge about what is out there will ever allow us to reach definite conclusions.

Questions about truth and its justification have also concerned statisticians who examined these questions from a normative perspective, focusing on the degree of confidence in conclusions that are based on empirical observations. These questions have been important in many applied areas as well, such as jury decisions and medical diagnosis (Dunning et al. 2004). In addition, in many real-life situations, confidence in one's judgements determines the likelihood of translating these judgements to action (Tversky and Koehler 1994; Koriat and Goldsmith 1996; Dunning 2007).

In experimental research, assessments of subjective confidence in one's own knowledge and judgements have been investigated over many years in a wide range of domains. These include perception and psychophysics, memory and metacognition, judgement and decision-making, and eyewitness testimony. Increased interest in confidence judgements can also be seen in such areas as social cognition, animal cognition, and neuroscience.

Two general issues have been addressed by researchers: the accuracy of metacognitive judgements and the bases of these judgements. With regard to the accuracy of metacognitive judgements, the observation that has attracted the attention of researchers in metacognition is that participants are generally accurate in monitoring their knowledge: They can tell when they know and when they do not know; and can judge when they are right and when they are wrong. For example, when studying a list of items, participants can predict with some accuracy which items they will recall at test (Nelson and Dunlosky 1991). During recall too, people can predict with some success which of the unrecallable memory targets they will be able to recognize among

distracters (Koriat 1993). In addition, when they are asked to answer each of several questions, participants can generally discriminate between correct and wrong answers (Goldsmith and Koriat 2008). The ability to monitor one's own knowledge was seen by Tulving and Madigan (1970) as 'one of the truly unique characteristics of human memory' (p. 477). This ability raises the question: how do people know that they know?

To answer this question, we must first examine the bases of metacognitive judgements. Understanding the bases of one's metacognitive judgements may provide a clue to both the accuracy and inaccuracy of people's knowledge of their own knowledge.

The bases of metacognitive judgements

Three general approaches to the bases of metacognitive judgements may be distinguished: the *direct-access approach*, the *information-based approach*, and the *experience-based approach* (see Koriat 2007). The direct-access view is perhaps best represented in the philosophy of knowledge, by the claims of rationalist philosophers that a priori truths (e.g. mathematical propositions) are based on intuition and deduction, and that their certainty is self-evident. In memory research, the direct-access approach assumes that metacognitive judgements are based on people's privileged access to the presence and strength of stored memory traces (see Dunlosky and Metcalfe 2009). For example, it was proposed that judgements of learning (JOLs) are based on detecting the strength of the memory trace that is formed following learning (e.g. Cohen et al. 1991). Similarly, tip-of-the-tongue (TOT) and feeling-of-knowing (FOK) judgements were claimed to monitor the actual presence of the elusive target in the memory store (Hart 1965; Burke et al. 1991; Yaniv and Meyer 1987). In the case of confidence judgements too, a direct access view generally underlies the use of such judgements in the context of strength theories of memory (see Van Zandt 2000).

In contrast to the direct-access view, a cue-utilization view has been gaining popularity in metacognition research (see Koriat 1997). According to this view, metacognitive judgements are inferential in nature, relying on a variety of beliefs and heuristics. A distinction is drawn, however, between information-based and experience-based judgements (Koriat et al. 2008). In information-based approaches, metacognitive judgements are assumed to rely on an analytic inference in which various considerations retrieved from long-term memory are consulted and weighed to reach an educated metacognitive judgement. For example, JOLs have been claimed to rely on the person's theories about how various characteristics of the study material or the conditions of learning influence memory performance (Koriat 1997; Benjamin 2003). Learners may also rely on their beliefs about their own skills and competence (Bandura 1997). Similarly, FOK judgements have been said to rest on deliberate inferences from one's own beliefs and knowledge (Nelson et al. 1984; Costermans et al. 1992). Discussions of subjective confidence also emphasize information-driven processes: confidence in two-alternative forced-choice (2AFC) general-knowledge question was claimed to rest on the balance of evidence in favour of the two answers (e.g. Koriat et al. 1980; Griffin and Tversky 1992; McKenzie 1997).

Unlike information-based approaches, which emphasize the content of domain-specific beliefs and knowledge retrieved from memory, experience-based approaches focus on the contribution of mnemonic cues that derive on-line from task performance. These cues are assumed to give rise automatically and unconsciously to a sheer metacognitive feeling (Koriat 2000; see Proust 2007, for a philosophical discussion). Indeed, extensive research has testified to the effects of internal cues on a variety of metacognitive judgements. Results suggest that JOLs made during study rest on the ease with which to-be remembered items are encoded or retrieved during learning (Nelson et al. 2004; Koriat and Ma'ayan 2005; Koriat et al. 2006; Karpicke 2009). FOK judgements have

been claimed to rely on the familiarity of the pointer that serves to probe memory (Reder 1988; Schwartz and Metcalfe 1992), or on the amount of partial clues that come to mind during the search for the memory target, and the ease with which they come to mind (Koriat 1993, 1995).

Confidence judgements seem also to rest on the fluency of selecting or retrieving an answer. Of particular relevance to the present work are findings indicating that participants express stronger confidence in the answers that they retrieve more quickly, whether those answers are correct or not (e.g. Kelley and Lindsay 1993; Robinson et al. 1997; Koriat et al. 2006). Largely, however, response speed is diagnostic of the correctness of the answer, so that the accuracy of confidence judgements is mediated in part by reliance on response latency (Costermans et al. 1992; Koriat and Ackerman 2010).

The processes underlying confidence judgements

Using the distinction between the three bases of metacognitive judgements, I will now outline several propositions regarding the processes underlying confidence judgements. To illustrate some of these propositions, I will use several informal observations regarding the reasons that people use to support some of the beliefs that they hold with strong conviction. For example, I would ask a student: 'What is your name?'. I would then ask: 'How confident are you that this is indeed your name?'. Generally, after an initial embarrassment, the answer is: 'Of course, one hundred percent'. When I then ask 'Why are you so confident?' the student would typically pause, and sometimes the immediate response is 'I just know'. Some students simply insist on a 'just know' response, perhaps implying a direct-access basis. Others venture to provide reasons, and these reasons seem often quite weak ('I remember that my girlfriend calls me Daniel. Actually she calls me Danny, but you know that Danny and Daniel are the same'; 'I can see my name printed on my driver's licence', etc.). Are these indeed the actual bases of one's strong conviction in one's own name? These and similar observations can help illustrate the following propositions regarding confidence judgements:

1. I propose that, in general, the *immediate* bases of feelings of confidence, as well as of other metacognitive feelings, lie primarily in mnemonic cues that derive from task performance rather than in the content of domain-specific declarative information retrieved from long-term memory. This proposal is based on observations in metacognition, which suggest that participants hardly apply their declarative knowledge and theories in making metacognitive judgements.

   For example, Koriat et al. (2004) found that JOLs made during learning were entirely indifferent to the expected retention interval, although actual recall exhibited the typical forgetting function. Thus, participants gave similar recall predictions whether they expected to be tested immediately after study, after a week, or even after a year. Koriat et al. proposed that JOLs rely primarily on encoding fluency, and that the fluency with which an item is encoded during study is not affected by when testing is expected. In addition, Kornell and Bjork (2009) found that JOLs fail to take into account the effects of number of study trials on memory (see also Kornell 2011; Kornell et al. 2011). Thus, learners do not apply spontaneously some of the most basic beliefs about learning and remembering in making recall predictions. They do so only under some specific conditions. For example, in Koriat et al.'s study, participants exhibited sensitivity to retention interval when they were asked to predict forgetting ('how many words will you forget') rather than remembering ('how many words will you recall'; see also Finn 2008).

   Furthermore, Koriat et al. (2008) had participants choose an answer to general-information questions, list reasons in support of their choice, and then indicate their confidence in the

correctness of the answer. When participants were required to list four supporting reasons, their confidence was lower than when they were required to list only one supporting reason. Thus, the effects of ease of retrieval (four reasons are more difficult to retrieve than one reason) can override the effects of the declarative content of the supporting reasons in affecting confidence judgements (Jacoby et al. 1989).

2. Information-driven processes, however, do play an important role in choice and confidence. It is proposed that when participants are presented with a 2AFC general-knowledge question, they engage typically in an analytic-like process, retrieving information from memory, and evaluating its implications before choosing the answer (see Koriat et al. 1980; Gigerenzer et al. 1991; Shafir et al. 1993). Often the pieces of information that come to mind consist of associations, hunches, and images that are not readily expressed in the form of declarative statements, but they can nevertheless tip the balance in one direction or the other. When participants have then to assess their confidence in their choice, they do not go over the entire protocol underlying their decision but rely primarily on the 'gist' of that protocol. They base their confidence on contentless mnemonic cues, such as the amount of deliberation and conflict that they had experienced in reaching the decision, and the speed with which the decision had been reached. These non-analytic cues (see Jacoby and Brooks 1984) represent the feedback from the *process* underlying the decision. Although these cues differ in quality from the considerations that were made in making the decision, they mirror significant aspects of the process that had determined the decision itself, primarily the balance of evidence in favour of the two options.

   As an analogy, we can think of a decision-making body that selects one of two alternatives based on majority rule. Once all the arguments have been heard and a vote has been cast, this vote is what finally matters. Likewise, confidence judgements would seem to rely primarily on the final vote—the overall impression formed after a deliberation regarding the relative support for each alternative. This overall impression is reflected in immediately available mnemonic cues, such as the amount of time it took to reach the decision. Perhaps, then, people are convinced about their own names not so much because of the content of individual considerations, but because of the 'unanimous vote'—the consensus among the variety of pieces of information that come to mind, and the ease and persistence with which they come to mind. Thus, it is proposed that as participants move from choosing an answer to assessing their confidence in that answer, the contribution of information-driven processes decreases and that of mnemonic cues increases.

3. The accuracy of metacognitive judgements depends largely on the extent to which the considerations and associations that come to mind lean towards the correct answer. Because these considerations and associations reflect the effects of learning and experience, they tend to support the correct answer. Proponents of the ecological probability approach (Brunswik 1956; Gigerenzer et al. 1991; Juslin and Olsson 1997; Fiedler 2007) have stressed the idea that people internalize the associations between cues and events in the world, and use the internalized knowledge when making metacognitive judgements. It is important to add that learning not only makes available declarative knowledge but also helps educate subjective experience itself. Information that is better learned, tends to be more readily retrievable, and tends to come to mind with greater consistency and persistence (Benjamin and Bjork 1996). Indeed, in a large number of studies, primitive subjective attributes, such as recognition, familiarity, fluency, and accessibility have been shown to provide valuable diagnostic information that can be used by the person as a basis for judgements (e.g. Kelley and Lindsay 1993; Koriat 1993; Goldstein and Gigerenzer 2002; Hertwig et al. 2008).

4. The processes underlying mnemonic-based metacognitive judgements occur largely outside of awareness (Proust 2008). This assumption contrasts with the spirit of information-based accounts of metacognitive judgements. For example, according to the theory of Probabilistic Mental Models (PMM; Gigerenzer et al. 1991) people choose between two answers by retrieving a cue that discriminates between the two answers. Associated with each cue is also a cue validity that describes how well that cue predicts the criterion. When the cue determines the choice, its cue validity is then reported as the confidence in the choice.

   Experience-based approaches, in contrast, assume that the process is much less analytic, and that people have little awareness of the mnemonic cues underlying their metacognitive judgements, let alone their cue validity (see Koriat et al. 2009). For example, in the mere exposure effect, repeated exposure to stimuli, even under subliminal presentation, has been found to lead to increased liking of these stimuli, although during debriefing, most participants predict that repeated exposures would lead to boredom and decreased liking (Murphy et al. 1995).

   Because metacognitive feelings rest on unconscious inferences (Jacoby et al. 1989), the phenomenology of these feelings is most consistent with the direct-access view. Metacognitive feelings often have the quality of direct perceptions (Kahneman 2003; Kahneman and Frederick 2005). A person in a TOT state, for example, can 'sense' the elusive name or word and can monitor its emergence into consciousness (Brown and McNeill 1966; see Schwartz and Metcalfe 2011). Subjective convictions in beliefs also have the quality of direct access. Therefore, the validity of metacognitive feelings is sometimes taken for granted by the person (Epstein and Pacini 1999), although such feelings may prove illusory in retrospect (Koriat 1994; Schwartz 1998). It would seem that direct-access accounts of metacognitive feelings derive their power primarily from the phenomenology of these feelings and from their general accuracy in predicting memory performance.
5. Because the heuristics that underlie immediate metacognitive feelings operate below full consciousness (Koriat 2000), when participants are asked to explain the reasons for their metacognitive feelings, they usually refer to declarative knowledge and theories rather than to the underlying mnemonic cues that derive from task performance. Never have I heard a participant justify his or her high JOL, FOK, or confidence by referring to such factors as processing fluency or ease of retrieval. Of course, the reasons mentioned by participants to justify their metacognitive feelings often capture some of the distal ecological influences that have shaped the mnemonic cues underlying these feelings. Going back to the conviction in one's own name, it is my argument, as I noted, that the student is convinced of his name because of the simple fact that every way he thinks about his name, the same name comes consistently, insistently and quickly to mind. However, the justifications mentioned by him may reflect the historical factors that are responsible for the mnemonic qualities associated with retrieving one's name. These qualities derive from one's own experience, such as the frequent usage of the name by one's acquaintances, the many instances in which one has to say or write one's name, and so forth.

In sum, the three approaches to the basis of metacognitive judgements may reflect different aspects of the processes underlying these judgements. Although these approaches imply qualitatively different processes, there is a great deal of overlap between their predictions. The mnemonic cues assumed to underlie subjective confidence mirror the information-based cues that drive the choice of an answer. In turn, the phenomenological quality of subjective convictions is seen to derive from the unconscious nature of mnemonic-based feelings, resulting in retrospective justifications of these feelings that stress declarative semantic and episodic considerations.

These propositions depart from what might be concluded from the preponderance of experimental findings demonstrating misleading effects of mnemonic cues (e.g. Chandler 1994; Koriat 1995; Benjamin et al. 1998; Brewer and Sampaio 2006). These demonstrations, which were intended to bring to the fore the contribution of mnemonic cues, have resulted in overemphasis on situations in which mnemonic cues drive judgements away from what would be implied by analytic considerations, resulting in faulty judgements. Under natural conditions, however, mnemonic cues tend to be valid, and their validity derives from the effects of learning and past experience.

Subjective confidence: the motivation for the present proposal

I will now describe some of the work that has led to the self-consistency model of subjective confidence. Some of the tasks that I used to study subjective confidence were intended to tap 'intuitive' judgements. These tasks were inspired by the idea of some philosophers that universally shared notions that are grasped by intuition, have the quality of self-evidence: they strike you as being right. One such task that I used was based on the well-known demonstration by Köhler (1947): 'There is a language that has names for different shapes. Guess which of these shapes is called Maluma and which is Takete'. Two observations were noteworthy: first, practically all participants matched the rounded shape with Maluma, but when I asked them to state the reasons for their choice, their reasons differed greatly across participants. Second, all participants expressed strong convictions in the correctness of their response to the extent that when I told some participants that they were wrong, the typical reaction was 'that's impossible!'. This is similar to the phenomenal feeling that philosophers associate with a priori or analytic truths: Such truths feel *necessarily* correct.

In the Maluma–Takete example, there is no right or wrong answer. However, similar observations were made with similar tasks in which there was a correct answer. One such task required the matching of antonymic words from non-cognate languages (e.g. *tuun–luk*) with their English equivalents (*deep–shallow*). This task had been used by researchers to examine the idea that a universal sound-meaning symbolism has been incorporated in the formation of all languages, and people have an intuitive feel for it. I was interested to know whether correct matches tend to be endorsed with stronger confidence than wrong matches. In one study, (Koriat 1975) participants' matches were found to be significantly better than chance, averaging 58.1%. In addition, the percentage of correct matches increased steeply with confidence judgements, suggesting that participants were successful in monitoring the correctness of their matches. The latter result presented a puzzle. Neither the information-based approach nor the experience-based approach offers a hint regarding the cues that participants might use to monitor their knowledge. The finding is reminiscent of the direct-access view that rationalists posit with regard to a priori propositions that are accessed through intuition.

An important feature of the word-matching task is that no simple algorithm exists for determining whether the answer is correct or wrong. However, such is also the case in many memory tasks in which participants are successful in monitoring the correctness of their answers. Thus, perhaps, there is some general principle that underlies the accuracy of monitoring in a variety of tasks, including memory tasks and the word-matching task.

In attempting to uncover such a principle, I reasoned that perhaps the observation that participants' matches were largely accurate ('knowledge') creates a confounding for the assessment of the confidence–accuracy correlation ('metaknowledge'). Because the correct match is the one that is consensually endorsed, perhaps confidence judgements are correlated with the consensuality of the match rather than with its correctness. To examine this possibility I tried to dissociate

between correctness and consensuality by including many items for which participants are likely to agree on the *wrong* match (Koriat 1976). The results clearly indicated that confidence ratings correlated with the consensuality of the match rather than with its correctness: For consensually-correct (CC) items, for which most participants chose the correct answer, correct answers were endorsed with stronger confidence, whereas for consensually-wrong (CW) items it was the *wrong* answers that were associated with stronger confidence. The *consensuality principle*—that confidence is correlated with the consensuality of the answer rather than with its correctness—has been replicated since for several other tasks as will be detailed later.

The conclusion from these results is that when a representative sample of items is used, participants are successful in monitoring the correctness of their responses, but they do that *indirectly* by relying on some cues that are correlated with accuracy. These cues would seem to underlie the consensuality of the response—the extent to which it tends to be endorsed by the majority of people. Thus, what Tulving and Madigan (1970) regarded as a truly unique characteristic of human memory turns out to be an artefactual consequence of the fact that in virtually all studies that examined the confidence-accuracy correlation in memory tasks, the consensually endorsed answer is the correct answer. That is, the percentage of correct answers in 2AFC questions is practically always above 50%. Thus, metaknowledge accuracy and knowledge accuracy are intimately linked: metaknowledge is accurate as long as knowledge itself is accurate.

The consensuality principle was also confirmed for response latency. Previous studies had established that response speed is diagnostic of accuracy, being faster for correct than for wrong answers (Kelley and Lindsay 1993; Robinson et al. 1997; Koriat et al. 2006). However, we showed that this is true only for CC items, whereas for CW items the opposite relationship is found (Koriat 2008, 2012).

The consensuality principle is a descriptive principle that does not offer a process account of the basis of confidence judgements and their accuracy. However, it may provide a lead to the question of how we know that we know. It suggests that what makes a person confident in a particular answer is what makes most people favour that answer in the first place. This idea motivated the development of the self-consistency model of subjective confidence (Koriat 2011, 2012; Koriat and Adiv 2011). Before describing the model, I would like to spell out its underlying metatheoretical assumptions.

A preamble to the model: philosophical perspectives

In this section, I would like to place the present proposal with respect to two major issues in epistemology. The first concerns the distinction between the rationalist and empiricist positions regarding the origin of knowledge, and the second concerns the distinction between correspondence and coherence theories of truth.

The origin of knowledge: rationalism versus empiricism

A central issue in the philosophy of knowledge is associated with the traditional distinction between rationalism and empiricism (see Edwards 1996; Markie 2008). The rationalist approach focuses on intuitive knowledge—a priori propositions whose truth is self-evident. Rationalists, such as Descartes, Spinoza, and Leibniz, maintained that there are significant aspects of our concepts and knowledge that are gained independent of sense experience. These are knowable either by direct intuition, or by deduction from intuited propositions. The examples mentioned include mathematical propositions, logical arguments, ethical or moral propositions, and even metaphysical beliefs (e.g. that God exists). Some rationalists posit that such truths are innate. Carruthers (1992), for example, argued that knowledge of some of the principles of folk-psychology (e.g. that

pain tends to be caused by injury) is innate. Innateness generally implies universality. As I noted earlier, my early research on the ability to guess the meaning of foreign words (Koriat 1975) was inspired by the notion that intuited, universal truths are phenomenologically experienced as self-evident.

In contrast, empiricists such as Locke and Berkeley argued that the origin of knowledge resides in the external world. According to them, sense experience is the ultimate source of knowledge and therefore the focus should be on a posteriori propositions whose justification relies on empirical observations.

Most philosophers, however, admit both sources of knowledge. Albert Einstein discussed the 'eternal antithesis between the two inseparable components of our knowledge, the empirical and the rational':

> We reverence ancient Greece as the cradle of western science. Here for the first time the world witnessed the miracle of a logical system which proceeded from step to step with such precision that every single one of its propositions was absolutely indubitable. I refer to Euclid's geometry. This admirable triumph of reasoning gave the human intellect the necessary confidence in itself for its subsequent achievements ….
>
> But before mankind could be ripe for a science which takes in the whole of reality, a second fundamental truth was needed, which only became common property among philosophers with the advent of Kepler and Galileo. Pure logical thinking cannot yield us any knowledge of the empirical world; all knowledge of reality starts from experience and ends in it. Propositions arrived at by purely logical means are completely empty as regards reality. Because Galileo saw this, and particularly because he drummed it into the scientific world, he is the father of modern physics – indeed, of modern science altogether. (Einstein 1934/1954, p. 271.)

At the risk of oversimplification, I would like to stress two aspects that distinguish the two types of knowledge. First, for rationalists, truth lies within: it can be grasped through 'pure reason'. In a sense, its acquisition is based on direct access (or else on a deduction from directly accessed truths). For empiricists, in contrast, knowledge originates from the outside world and hence ultimately relies on empirical observations. Second, there is a consensus that intuition and deduction provide beliefs whose truth is self-evident, and is beyond any doubt ('absolutely indubitable' in Einstein's words). These beliefs are endowed with a sense of necessity. Empiricists, in contrast, admit a degree of uncertainty, arguing, for example, that we can never be sure that our sensory impressions are true.

These comments suggest that perhaps different processes underlie confidence judgements when knowledge originates from within than when it originates from without. Taken together, however, the results of Koriat (1975, 1976, 2008, 2011) suggest otherwise. Furthermore, with regard to intuitive knowledge, the extensive work on intuitive feelings by experimental psychologists (see Lieberman 2000; Hogarth 2001; Myers 2002; Kahneman 2003; Plessner et al. 2007) raises concern about the assumptions among some philosophers that there exists an intimate link between intuition and a priori, innate knowledge, and that intuition provides knowledge whose truth is absolutely certain. Not only has there been evidence that intuitive, gut feelings can have their origin in experience (Westcott 1968; Reber 1989), but also that intuitive feelings that are held with strong subjective certainty are sometimes wrong (Denes-Raj and Epstein 1994; Koriat 1994, 1998; see Nagel 2007). This evidence blurs the distinction between knowledge originating from within and knowledge originating from without, and invites a common framework in which subjective confidence in both types of knowledge can be analysed.

To build such a framework, consider the psychological situation of a participant who is required to assess the confidence in the answer to such questions as 'Which city has more inhabitants, Hanover or Bielefeld?' (Gigerenzer et al. 1991), or 'What is the capital of Australia, Canberra or

Sydney?' (Fischhoff et al. 1977). The pertinent clues for the answer must be retrieved from one's own memory rather than (directly) from the outside world. In this respect, the situation is not different from that underlying the verification of analytic truths. Such is also the case when the propositions concern semantic knowledge, episodic memory, or social and metaphysical beliefs (e.g. 'There is a supreme being controlling the universe'; see later). In attempting to validate one's memories or beliefs (see Ross 1997) or to judge the source of one's memories (see Mitchell and Johnson 2000; Lindsay 2008) one must make do with a variety of pieces of information accessed from within. Indeed, a recent functional magnetic resonance imaging study suggests that the neural activity related to metacognitive judgements is characterized by a shift away from externally directed cognition toward internally directed cognition (Chua et al. 2009). So what is the basis of one's degree of certainty in an answer that is retrieved from memory?

The first postulate underlying SCM is that although the validation of one's own knowledge is based on retrieving information from memory, the underlying process is analogous to that in which information is sampled from the outside world with the goal (1) to test a hypothesis about a population, and (2) to assess the likelihood that the conclusion reached is correct. I argue that such is the case whether participants need to validate propositions whose truth is a priori or propositions whose truth is a posteriori. Thus, the prototype for the underlying process is provided by the statistical procedures that are used by researchers in attempting to draw conclusions about the external world: a proximal sample of observations is used to make inferences about some 'true' parameter of a distal population. The critical difference, of course, is that information is sampled from within rather than from without. The model to be sketched as follows incorporates this assumption.

Correspondence versus coherence theories of truth

I turn now to the second issue, which helps introduce the second postulate. This issue concerns the distinction between two major philosophical theories of truth, correspondence theories and coherence theories (Kirkham 1992). Correspondence theories posit that the truth or falsity of a statement is determined only by how it relates to the world, and whether it accurately describes objects or facts. Coherence theories, in contrast, assume that the truth or falsity of a statement is determined by its relations to other statements rather than its relation to the world (Rescher 1973; Walker 1989). In this view, a person's belief is true if it is coherent with his or her body of beliefs, that is, if it is a constituent of a systematically coherent whole.

The correspondence view of truth reflects the intentions of confidence judgements. Confidence in a proposition reflects the likelihood that that proposition agrees with reality (e.g. that one's name is indeed Daniel, or that Canberra is indeed the capital of Australia). The problem, however, is how can one assess such agreement with reality if one does not have access to reality independent of what one knows or believes about it? Kant stated this problem as follows:

> Truth is said to consist in the agreement of knowledge with the object. According to this mere verbal definition, then, my knowledge, in order to be true, must agree with the object. Now, I can only compare the object with my knowledge by this means, namely, by taking knowledge of it. My knowledge, then, is to be verified by itself, which is far from being sufficient for truth. For as the object is external to me, and the knowledge is in me, I can only judge whether my knowledge of the object agrees with my knowledge of the object. Such a circle in explanation was called by the ancients Diallelos. (Kant 1885, p. 40.)

The resolution of this issue calls for a second postulate: although confidence judgements pertain to correspondence, the mnemonic cue for metacognitive assessments of correspondence is degree of coherence. Confidence in an answer or belief depends on the extent to which that

answer or belief is supported consistently by the various pieces of information that come to mind. Indeed, several discussions have stressed the use of internal consistency as a cue for the validity of one's own beliefs. Ross (1997), for example, noted that people rely on internal coherence in judging the validity of their recollections, and use incoherence and internal contradictions as good reasons to doubt the reality of recollections.

In epistemological discussions, the notion of coherence has been discussed extensively in connection with the justification of beliefs. Unlike *foundationalist* theories, which assume that beliefs are justified on the basis of other beliefs, *Coherentism* theories claim that a belief is justified by the way it fits together with the rest of the belief system of which it is a part (BonJour 1985). Foundationalists escape the regress problem of an infinite chain of justification (see Moser 1988) by postulating the existence of justified basic beliefs that do not owe their justification to other beliefs (Van Cleve 2005). Coherentists, in contrast, avoid the regress problem without postulating the existence of non-inferential basic beliefs.

The notion of coherence that I assume to underlie confidence judgements is quite loose. First, I assume that what matters is only the internal consistency within the set of thoughts that are activated during the attempt to answer a question or validate a belief. In this respect, coherence or consistency can be said to be output-bound (Koriat and Goldsmith 1996), relative to the set of clues that are activated. Second, what is activated during the choice of an answer is generally an assortment of images, memories, beliefs, associations, and thoughts that cannot always be expressed in a propositional form. Therefore, coherence reflects the extent to which these clues produce a sense of convergence versus a sense of tension or conflict. Indeed, studies of the illusory-truth effect indicate that mere familiarity and fluency can enhance truth judgements. For example, the repetition of a statement increases its perceived truth even when the statements are actually false (Hasher et al. 1977; Bacon 1979; Arkes et al. 1989). Truth judgements are also enhanced by perceptual fluency (e.g. visual contrast; Reber and Schwarz 1999; Hansen et al. 2008; Unkelbach and Stahl 2009) and by manipulations that increase contextual fluency (placing the statement in contexts that provide a continuity of meaning; Parks and Toth 2006).

In sum, because people have no access to the object of their beliefs over and above what they know about it, they rely on a fast assessment of overall coherence (see Bolte and Goschke 2005) as a basis for their judgements about correspondence. In terms of Polanyi's (1958) terminology, the 'object' of metacognitive judgements is correspondence, but the 'tool' is coherence. This state of affairs raises a dilemma for the evaluation of the accuracy of one's confidence judgements: should these judgements be evaluated against correspondence, because this is what participants feel (and state), or should they be evaluated against coherence? As we note later, the discrepancy between the two criteria may explain the overconfidence bias observed in calibration research.

The self-consistency model of subjective confidence rests on the two postulates mentioned earlier. First, it assumes that although information is retrieved from memory, the process is similar to the statistical procedure involved in assessing confidence in a sample-based inference about the outside world. Second, coherence or reliability is used as a cue for validity.

The self-consistency model of subjective confidence

I will now present briefly the SCM of subjective confidence. Underlying SCM is a metaphor of the person as an intuitive statistician (Peterson and Beach 1967; Gigerenzer and Murray 1987; see McKenzie 2005). People's confidence judgements are modelled by the classical procedures of calculating statistical level of confidence when conclusions about a population are to be made based on a sample of observations. When faced with a 2AFC general-information question, or a question about some social or metaphysical belief, it is by replicating the choice process several

times that a person can appreciate the degree of doubt or certainty involved. The assessment of degree of certainty is obtained by sampling different 'representations' or considerations from memory and assessing the extent to which they agree in favouring a particular decision. Subjective confidence essentially represents an assessment of *reproducibility*—the likelihood that a new sample of representations drawn from the same population will yield the same choice. Thus, reliability is used as a cue for validity.

SCM does not pretend to describe the complex processes involved in making a choice, but only to capture the *feedback* from that process. It is assumed that this feedback, which affects confidence, is a crude sense of consistency that can be modelled by a simple count of the proportion of representations favouring each of the two alternatives (see Alba and Marmorstein 1987). A detailed description of the model can be found elsewhere (Koriat 2011, 2012). Here only a brief description will be presented of a specific implementation of the model.

An important assumption of SCM is that in responding to 2AFC items, whether they involve general-information questions or beliefs and attitudes, participants with the same experience draw representations largely from the same, commonly shared population of representations associated with each item. If each representation favours one of the two answers, each item can be characterized by a probability distribution, with p_{maj} denoting the probability that a representation favouring the majority alternative will be sampled.

Given a particular value of n, the number of representations sampled, the parameter p_{maj} for a given item may be estimated from the probability with which the majority alternative is chosen. This probability can be indexed operationally by the proportion of participants who choose the preferred alternative ('item consensus'), or by the proportion of times that the same participant chooses the preferred alternative across repeated presentations ('item consistency'). For example, for an item with a 40–60% between-participant split of choices, item consensus will be 60%.

One version of the model assumes that participants sample a maximum of seven representations, each of which yields a binary subdecision, and that the overt choice is dictated by the majority vote. However, if three representations in a row yield the same subdecision, the search is stopped and the Run-3 subdecision is reported. An index of self-consistency was used, which is related to the standard deviation of the subdecisions: $1-\sqrt{\hat{p}\hat{q}}$. It is calculated over the actual number of representations sampled.

A simulation experiment that incorporates these simple assumptions yielded the results depicted in Fig. 13.1. These results indicate the functions relating the index of self-consistency to the probability of choosing the majority answer, pc_{maj} for majority and minority choices. Three features should be noted. First, mean self-consistency (and hence, confidence) for each item should increase with pc_{maj}. Second, self-consistency is systematically higher for majority than for minority choices. Finally, whereas for majority choices, self-consistency increases steeply with pc_{maj}, for minority choices, it decreases but much more shallowly.

Why is self-consistency lower for minority than for majority choices? The reason is that when a sample of representations happens to favour a minority choice, the proportion of subdecisions favouring that choice will be smaller on average than when the sample favours the majority choice. For example, for $p_{maj} = 0.70$, and $n = 7$, the likelihood that six or seven representations will favour the majority answer is 0.329, whereas only in 0.004 of the samples will six or seven representations favour the minority answer.

The simulation experiment mentioned earlier indicated that the results for n_{act}, the number of representations actually drawn, mimic very closely those obtained for self-consistency. Assuming that response latency increases with n_{act}, it should be longer for minority than for majority choices.

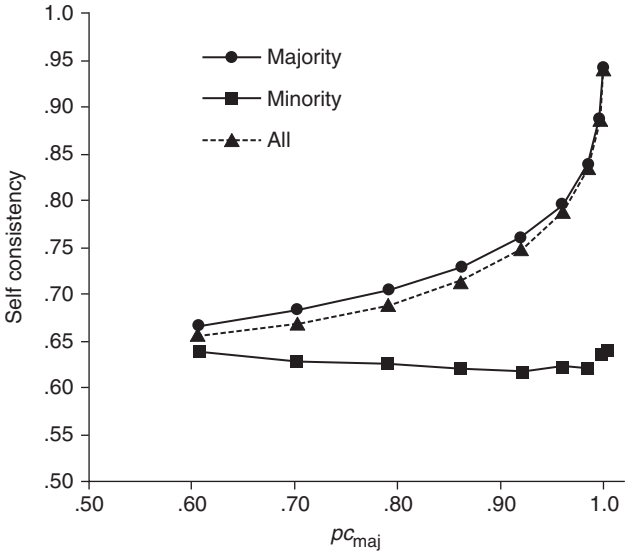


Fig. 13.1 Self-consistency scores as a function of the probability of choosing the majority option (PC_{maj}) based on the results of the simulation experiment. The results are plotted separately for majority and minority choices. Reprinted from figure 2, panel A, in 'The Construction of Attitudinal Judgments: Evidence from Attitude Certainty and Response Latency' by A. Koriat and S. Adiv, *Social Cognition*, 29, 2011, 587, Copyright 2011 by Guilford Press.

Note that the results in Fig. 13.1 were obtained under the assumption that participants choose the alternative that is favoured by the *majority* of representations in *their* accessed sample of representations. This pattern should be obtained both in a within-individual analysis and in a between-individual analysis.

Empirical evidence

I will present a brief summary of the results of several experiments in which these predictions were tested.

The relationship between confidence and cross-person consensus

As noted, pc_{maj} can be indexed by the proportion of participants who choose the majority answer. To test the predictions of the model, the answer that was chosen by the majority of participants for each item was designated ad hoc as the consensual (majority) answer, and the other as the non-consensual (minority) answer. Mean confidence was then plotted as a function of item consensus—the proportion of participants who chose the majority answer. This was done separately for majority and minority answers. The results yielded a pattern that is qualitatively similar to that depicted in Fig. 13.1. This was true across several tasks: general knowledge, word matching, comparison of the length of two lines, comparison of the area of two figures, social beliefs, and social attitudes (Koriat 2011, 2012; Koriat and Adiv 2011). In all of the tasks, participants made a two-alternative choice and expressed their confidence in the choice. The generality of the findings across domains supports the assumption of SCM that confidence is based on mnemonic cues that are indifferent to the specific content of the representations that are sampled.

The systematic difference between majority and minority choices can explain the consensuality principle: Participants are more confident when their choice is consistent with that of most participants. This should be the case even if all participants are assumed to draw their representations from a commonly shared population (see Fig. 13.1).

The relationship between confidence and within-person consistency

SCM was tested also in a within-individual design. Participants were presented repeatedly with the same set of 2AFC items. The answers to each item were then classified as frequent or rare depending on their relative frequency across repetitions, and confidence was plotted for the frequent and rare answers as a function of item consistency—the relative frequency of the frequent (majority) choice across repetitions.

The predictions of SCM for a within-person design were tested for general knowledge, word matching, perceptual judgements, and social beliefs and attitudes. In all of these domains, the results were in line with predictions. In particular, participants were more confident when they made their more frequent choice than when they made their less frequent choice.

Another result that was observed is that confidence in the first presentation predicted the likelihood of making the same choice in subsequent presentations of the item. This is consistent with the assumption that subjective confidence in a choice monitors reproducibility—the likelihood of making the same choice in a subsequent presentation of the item.

Response latency

All of the results summarized so far were replicated when response speed rather than confidence was used as the dependent variable. Thus, response latency was overall shorter for consensual choices than for non-consensual choices and for frequent choices than for rare choices. Overall, the results suggest that response speed is a frugal cue for self-consistency and can be used as a basis for confidence. The results also indicated that the speed of a choice predicts the reproducibility of the choice.

The correlation between confidence and accuracy

The results for the confidence–accuracy correlation also yielded clear support for the consensuality principle. This was true for general-information questions (Koriat 2008), FOK judgements (Koriat 1995), and perceptual judgements (Koriat 2011). It was also observed for sentence memory (Brewer and Sampaio 2006). Both confidence and response speed were correlated with the consensuality of the choice rather than with its correctness: The confidence–accuracy correlation was positive when the consensual choice was the correct choice but negative when it was the wrong choice.

These results disclose the link between knowledge and metaknowledge (Koriat 1993): people know that they know because (or when) they know. Indeed, for the CW items people are 'doubly cursed' (Dunning et al. 2003): they do not know, and do not know that they do not know.

The results for perceptual judgements (Koriat 2011) also supported the consistency principle, which is analogous to the consensuality principle: The confidence-accuracy correlation was positive for items in which the participant's frequent choice was the correct choice but negative for items in which the frequent choice was the wrong choice. These results were also mimicked by the results for response speed.

The calibration of confidence judgements

SCM also provides an account of the overconfidence bias that has been observed in calibration studies (Lichtenstein et al. 1982; Griffin and Brenner 2004). According to SCM, the overconfidence

bias derives, in part, from participants' reliance on reliability as a cue for validity. Reliability (or consistency) is practically always higher than validity. Confidence judgements are assumed to monitor self-consistency but their accuracy is evaluated in calibration studies against correctness. Indeed, the overconfidence bias was reduced or eliminated when confidence was evaluated against indexes of self-consistency rather than against correctness (Koriat 2011).

Interparticipant consensus in choice and confidence

Consistent with SCM, all of the tasks mentioned exhibited a marked degree of cross-person consensus, suggesting that participants share the same core of item-specific representations from which they draw their sample of representations on each occasion. This was true even for social beliefs and social attitudes. Furthermore, cross-person consensus and within-person consistency were correlated so that the choices that evidenced higher within-person consistency were more likely to be made by other participants.

Discussion

The question of how we can be certain about our beliefs has intrigued philosophers for centuries, and has been addressed in a broad range of domains. Subjective confidence has also attracted much interest in view of the many observations testifying for serious deficiencies in the ability to monitor one's own knowledge and performance (Burton 2008).

In this chapter, I described briefly a model of subjective confidence, focusing on the metatheoretical assumptions underlying the model. In what follows, I discuss these assumptions. SCM assumes that confidence judgements are inferential in nature, relying primarily on cues that derive from task performance. This view departs from the direct access view, which assumes that metacognitive judgements are based on privileged access to memory traces. It also departs from the view that these judgements are mediated by an analytic process in which declarative propositions retrieved from long-term memory are consulted to reach an educated metacognitive assessment. Rather, confidence in a decision is parasitic on the process of making a decision, and is based on mnemonic cues that derive online from that process (Koriat et al. 2008).

As noted in the introduction, one of the central issues in philosophy concerns the origin of knowledge. For rationalist philosophers, the origin of knowledge lies within the person whereas for empiricist philosophers it lies without. However, it was argued that in a typical situation in which participants are required to validate a proposition, they must draw on information that resides within, whether that proposition concerns semantic and episodic memory or so-called a priori truth. Therefore, it was proposed that the self-consistency model might apply not only to memory questions that depend on real-world knowledge but also to statements concerning personal and metaphysical beliefs.

Although the clues for confidence must come from within, it was argued that the process has much in common with that in which information is retrieved from without. Specifically, in testing a hypothesis about a population based on a sample of observations, researchers generally put greater trust in the hypothesis as a function of the level of significance with which the null hypothesis is rejected. That is, they behave as if the correctness of the hypothesis, as well as the likely reproducibility of the observed result, is a monotonically increasing function of level of confidence (see Schervish 1996; Dienes 2011). Statistical level of confidence increases with decreased variance—the extent to which the sampled observations consistently support the hypothesis. Let us examine this idea closely as it bears on the distinction between coherence and correspondence.

Assume that we wish to test the hypothesis that among married couples, husbands are happier than their wives. We draw randomly one couple from a population and find that indeed the

husband is happier than his wife. Apart from the fact that most people will not put too much faith in a conclusion that is based on a sample of $n = 1$, the problem is that such a sample does not allow assessment of the credibility of the conclusion.

The situation changes radically when a larger sample is drawn, e.g. $n = 100$. In this case, statistical level of confidence is based on the *internal consistency* within the sample. If we find that in 80 of the 100 couples husbands are happier, our faith in the hypothesis that in the 'real world' (i.e. in the population as a whole) husbands tend to be happier stems from the consistency *within the sample*. Thus, in a sense, coherence is used as a cue for correspondence.

SCM assumes that in a similar manner, it is by replicating the choice process several times that people appreciate the amount of doubt involved. Subjective confidence in the *validity* of a proposition is then based on the *reliability* with which the proposition is supported across the sample of representations.

This view differs from that of the PMM theory, which assumes that confidence is based on the stored validity of a single cue that discriminates between the two alternative answers. Clearly, PMM is an inferential, cue-based model as far as the choice of an answer is concerned. However, when it comes to confidence, the model is more like a trace-access model because confidence is read out directly from the stored validity of the cue.

Unlike PMM theory, SCM assumes that confidence depends on the internal consistency within a *collection* of representations. This assumption avoids the regress problem without postulating a direct-access basis for confidence. The logic underlying SCM is the same as that underlying the (mis)interpretation of statistical level of confidence as capturing the degree of trust in a hypothesis. The finding that confidence in the first presentation of an item predicts the likelihood of making the same choice in subsequent presentations of the item also parallels the (mis)interpretation of statistical level of confidence as capturing the likely reproducibility of the observed effect.

In line with SCM, confidence judgements were found to track both the stable and variable contributions to choice. The stable contributions stem from the constraints imposed by the population of representations available in memory. In general, the polarization of the population of representations associated with an item constraints the extent of fluctuation in judgements that may be expected across occasions and across people. The variable contributions are disclosed by the systematic differences between majority and minority choices, which are assumed to convey information about the specific sample of representations underlying a particular choice (Koriat and Adiv 2011; see also Wright 2010).

The finding that the same pattern of results was obtained across different domains reinforces the assumption that confidence is based on structural, contentless cues. This finding may also be taken to imply that from a psychological point of view the processes underlying confidence in a priori truths are not qualitatively different from those underlying confidence in a posteriori truths. Admittedly, in the case of a priori truths (e.g. that the internal angles of a triangle add up to 180 degrees or that two plus two equals four), there is generally little variance between the outcomes of different representations of the question. However, perhaps that is precisely the cue for the strong conviction associated with such statements: What characterizes a priori beliefs is that however one thinks of them one arrives at the same conclusion. Nevertheless, the question should be entertained whether there are particular beliefs for which we should postulate some sort of direct access.

This question actually applies to episodic knowledge as well, when such knowledge is held with strong confidence (e.g. one's name). Metcalfe (2000), for example, postulated a 'special noetic state' in which metacognitive judgements are based on direct access rather than on inference from cues. Unkelbach and Stahl (2009) also proposed that when judging the truth of a statement,

'participants may simply know the factual truth or falsity of a statement and judge it accordingly' (p. 24). Gigerenzer et al.'s PMM model (1991) also incorporates a strategy (*local mental model*) in which a choice of an answer is based on a direct solution by memory. Only when this strategy fails, do participants construct a PMM that uses probabilistic information from a natural environment. Thus, an important question that we leave open is whether there are beliefs for which subjective confidence depends on a process that is qualitatively different from that postulated by SCM.

Acknowledgement

I am grateful to Shiri Adiv for her help and advice in the preparation of the manuscript and to Rinat Gil and Dana Klein for their help in the statistical analyses.

References

Alba, J. W. and Marmorstein, H. (1987). The effects of frequency knowledge on consumer decision making. *Journal of Consumer Research*, 14, 14–25.

Arkes, H. R., Hackett, C., and Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, 2, 81–94.

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning & Memory*, 5, 241–52.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, 31, 297–305.

Benjamin, A. S. and Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.) *Implicit memory and metacognition*, pp. 309–38. Mahwah, NJ: Erlbaum.

Benjamin, A. S., Bjork, R. A., and Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, 127, 55–68.

Bolte, A. and Goschke, T. (2005). On the speed of intuition: Intuitive judgments of semantic coherence under different response deadlines. *Memory & Cognition*, 33, 1248–55.

BonJour, L. (1985). *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.

Brewer, W. F. and Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory*, 14, 540–52.

Brown, R. and McNeill, D. (1966). The 'tip of the tongue' phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5, 325–37.

Brunswik, E. (1956). *Perception and representative design in psychological experiments*. Berkeley, CA: University of California Press.

Burke, D. M., MacKay, D. G., Worthley, J. S., and Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, 30, 542–79.

Burnet, J. (1930). *Early Greek philosophy*. London: Adam and Charles Black.

Burton, R. A. (2008). *On being certain: Believing you are right even when you're not*. New York: St. Martin's Press.

Carruthers, P. (1992). *Human knowledge and human nature*. Oxford: Oxford University Press.

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, 22, 273–80.

Chua, E. F., Schacter, D. L., and Sperling, R. A. (2009). Neural correlates of metamemory: A comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of Cognitive Neuroscience*, 21, 1751–65.

Cohen, R. L., Sandler, S. P., and Keglevich, L. (1991). The failure of memory monitoring in a free recall task. *Canadian Journal of Psychology*, 45, 523–38.

Costermans, J., Lories, G., and Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 142–50.

Denes-Raj, V. and Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, 66, 819–29.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–90.

Dunlosky, J. and Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage.

Dunning, D. (2007). Prediction: The inside view. In A. W. Kruglanski and E. T. Higgins (Eds.) *Social Psychology: Handbook of basic principles*, pp. 69–90. New York: The Guilford Press.

Dunning, D., Johnson, K., Ehrlinger, J., and Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12, 83–7.

Dunning, D., Heath, C., and Suls, J. M. (2004). Flawed self-assessment. *Psychological Science*, 5, 69–106.

Edwards, P. (Ed.) (1996). *The encyclopedia of philosophy*, Vols. *1 and 2—complete and unabridged*. New York: Macmillan Reference.

Einstein, A. (1934/1954). *Ideas and opinions* (S. Bargmann trans.). New York: Crown Publishers.

Epstein, S. and Pacini, R. (1999). Some basic issues regarding dual-process theories from the perspective of cognitive-experiential self-theory. In S. Chaiken and Y. Trope (Eds.) *Dual process theories in social psychology*, pp. 462–82. New York: Guilford Press.

Fiedler, K. (2007). Information ecology and the explanation of social cognition and behavior. In A. W. Kruglanski and E. T. Higgins (Eds.) *Social psychology: Handbook of basic principles*, pp. 176–200. New York: The Guilford Press.

Finn., B. (2008). Framing effects on metacognitive monitoring and control. *Memory Cognition*, 36, 813–21.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 552–64.

Gigerenzer, G., and Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506–28.

Goldsmith, M. and Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. Benjamin and B. Ross (Eds.) *Psychology of learning and motivation, Vol. 48: Memory use as skilled cognition*, pp. 1–60. San Diego, CA: Elsevier.

Goldstein, D. G. and Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109, 75–90.

Griffin, D. and Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler and N. Harvey (Eds.) *Blackwell handbook of judgment and decision making*, pp. 177–99. Malden, MA: Blackwell Publishing.

Griffin, D. and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411–35.

Hansen, J., Dechêne, A., and Wänke, M. (2008). Discrepant fluency increases subjective truth. *Journal of Experimental Social Psychology*, 44, 687–91.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56, 208–16.

Hasher, L., Goldstein, D., and Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16, 107–12.

Hertwig, R., Herzog, S. M., Schooler, L. J., and Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1191–206.

Hogarth, R. (2001). *Educating intuition*. Chicago, IL: University of Chicago Press.

Jacoby, L. L. and Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. H. Bower (Ed.) *The psychology of learning and motivation: Advances in research and theory*, pp. 1–47. New York: Academic Press.

Jacoby, L. L., Kelley, C. M., and Dywan, J. (1989). Memory attributions. In H. L. Roediger and F. I. M. Craik (Eds.) *Varieties of memory and consciousness: Essays in honour of Endel Tulving*, pp. 391–422. Hillsdale, NJ: Erlbaum.

Juslin, P. and Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, 104, 344–66.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58, 697–720.

Kahneman, D. and Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak and R. G. Morrison (Eds.) *The Cambridge handbook of thinking and reasoning*, pp. 267–93. Cambridge: Cambridge University Press.

Kant, I. (1885). *Kant's introduction to logic, and his Essay on the mistaken subtility of the Four Figures* (T. K. Abbott, trans.) London: Longmans, Green and Co.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, 138, 469–86.

Kelley, C. M. and Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, 32, 1–24.

Kirkham, R. L. (1992). *Theories of truth: A critical introduction*. Cambridge, MA: MIT Press.

Köhler, W. (1947). *Gestalt psychology*. New York: The New American Library of World Literature, Inc.

Koriat, A. (1975). Phonetic symbolism and the feeling of knowing. *Memory & Cognition*, 3, 545–8.

Koriat, A. (1976). Another look at the relationship between phonetic symbolism and the feeling of knowing. *Memory & Cognition*, 4, 244–8.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609–39.

Koriat, A. (1994). Memory's knowledge of its own knowledge: The accessibility account of the feeling of knowing. In J. Metcalfe and A. P. Shimamura (Eds.) *Metacognition: Knowing about knowing*, pp. 115–35. Cambridge, MA: MIT Press.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General*, 124, 311–33.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126, 349–70.

Koriat, A. (1998). Illusions of knowing: The link between knowledge and metaknowledge. In V. Y. Yzerbyt, G. Lories, and B. Dardenne (Eds.) *Metacognition: Cognitive and social dimensions*, pp. 16–34. London: Sage.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149–71.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, and E. Thompson (Eds.) *The Cambridge handbook of consciousness*, pp. 289–325. New York: Cambridge University Press.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 945–59.

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, 140, 117–39.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, 119, 80–113.

Koriat, A. and Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science,* 13, 441–53.

Koriat, A. and Adiv, S. (2011). The construction of attitudinal judgments: Evidence from attitude certainty and response latency. *Social Cognition*, 29, 577–611.

Koriat, A. and Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, 103, 490–517.

Koriat, A. and Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, 52, 478–92.

Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 107–18.

Koriat, A., Bjork, R. A., Sheffer, L., and Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, 133, 643–56.

Koriat, A., Ma'ayan, H., and Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, 135, 36–69.

Koriat, A., Nussinson, R., Bless, H., and Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky and R. A. Bjork (Eds.) *Handbook of memory and metamemory*, pp. 117–35. New York: Psychology Press.

Koriat, A., Ackerman, R., Lockl, K., and Schneider, W. (2009). The memorizing-effort heuristic in judgments of learning: A developmental perspective. *Journal of Experimental Child Psychology*, 102, 265–79.

Kornell, N. (2011). Failing to predict future changes in memory: A stability bias yields long-term overconfidence. In A. S. Benjamin (Ed.) *Successful remembering and successful forgetting: a Festschrift in honor of Robert A. Bjork*, pp. 365–86. London: Psychology Press.

Kornell, N. and Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, 138, 449–68.

Kornell, N., Rhodes, M. G., Castel, A. D., and Tauber, S. K. (2011). The ease of processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, 22, 787–94.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases*, pp. 306–34. New York: Cambridge University Press.

Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin*, 126, 109–37.

Lindsay, D. S. (2008). Source Monitoring. In H. L. Roediger, III (Ed.) *Cognitive psychology of memory. Vol. 2 of Learning and Memory: A Comprehensive Reference* (J. Byrne, Ed.), pp. 325–48. Oxford: Elsevier

Markie, P. (2008). Rationalism vs. empiricism. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. Stanford, CA: The Center for the Study of Language and Information (CSLI), Stanford University.

McKenzie, C. R. M. (1997). Underweighting alternatives and overconfidence. *Organizational Behavior and Human Decision Processes*, 71, 141–60.

McKenzie, C. R. M. (2005). Judgment and decision making. In K. Lamberts and R. L. Goldstone (Eds.) *Handbook of cognition*, pp. 321–38. London: Sage.

Metcalfe, J. (2000). Feelings and judgments of knowing: Is there a special noetic state? *Consciousness and Cognition*, 9, 178–86.

Mitchell, K. J. and Johnson, M. K. (2000). Source monitoring: Attributing mental experiences. In E. Tulving and F. I. M. Craik (Eds.) *The Oxford handbook of memory*, pp. 179–95. New York: Oxford University Press.

Moser, P. K. (1988). Internalism and coherentism: A dilemma. *Analyses*, 48, 161–3.

Murphy, S. T., Monahan, J. L., and Zajonc, R. B. (1995). Additivity of nonconscious affect: Combined effects of priming and exposure. *Journal of Personality and Social Psychology*, 69, 589–602.

Myers, D. G. (2002). *Intuition: Its powers and perils*. New Haven, CT: Yale University Press.

Nagel, J. (2007). Epistemic intuitions. *Philosophy Compass*, 2, 792–819.

Nelson, T. O. and Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The 'delayed-JOL effect.' *Psychological Science*, 2, 267–70.

Nelson, T. O., Gerler, D., and Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General*, 113, 282–300.

Nelson, T. O., Narens, L., and Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods*, 9, 53–69.

Parks, C. M. and Toth, J. P. (2006). Fluency, familiarity, aging, and the illusion of truth. *Aging, Neuropsychology, and Cognition*, 13, 225–253.

Peterson, C. R. and Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68, 29–46.

Plessner, H., Betsch, C., and Betsch, T. (2007) (Eds.) *Intuition in judgment and decision making*. Mahwah, NJ: Erlbaum.

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. Chicago, IL: University of Chicago Press.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 159, 271–95.

Proust, J. (2008). Epistemic agency and metacognition: An externalist view. *Proceedings of the Aristotelian Society*, 108, 241–68.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118, 219–35.

Reber, R. and Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, 8, 338–42.

Reder, L. M. (1988). Strategic control of retrieval strategies. In G. H. Bower (Ed.) *The psychology of learning and motivation: Advances in research and theory*, *Vol. 22*, pp. 227–59. New York: Academic Press.

Rescher, N. (1973). *Coherence theory of truth*. London: Oxford University Press.

Robinson, M. D., Johnson, J. T., and Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology*, 82, 416–25.

Ross, M. (1997). Validating memories. In N. L. Stein, P. A. Ornstein, B. Tversky, and C. Brainerd (Eds.) *Memory for everyday and emotional events*, pp. 49–81. Mahwah, NJ: Erlbaum.

Schervish, M. J. (1996). *P* Values: What they are and what they are not. *The American Statistician*, 50, 203–6.

Schwartz, B. L. (1998). Illusory tip-of-the-tongue states. *Memory*, 6, 623–42.

Schwartz, B. L. and Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1074–83.

Schwartz, B. L. and Metcalfe, J. (2011). Tip-of-the-tongue (TOT) states: Retrieval, behavior, and experience. *Memory & Cognition*, 39, 737–49.

Shafir, E., Simonson, I., and Tversky, A. (1993). Reason-based choice. *Cognition*, 49, 11–36.

Tulving, E., and Madigan, S. A. (1970). Memory and verbal learning. *Annual Review of Psychology*, 21, 437–84.

Tversky, A. and Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547–67.

Unkelbach, C. and Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness & Cognition*, 18, 22–38.

Van Cleve, J. (2005). Why coherence is not enough: A defense of moderate foundationalism. In M. Steup and E. Sosa (Eds.) *Contemporary debates in epistemology*, pp. 168–80. Oxford: Blackwell.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 582–600.

Walker, R. C. S. (1989). *The coherence theory of truth: Realism, anti-realism, idealism*. London: Routledge.

Westcott, M. R. (1968). *Toward a contemporary psychology of intuition*. New York: Holt, Rinehart and Winston.

Wright, J. C. (2010). On intuitional stability: The clear, the strong, and the paradigmatic. *Cognition*, 115, 491–503.

Yaniv, I. and Meyer, D. E. (1987). Activation and metacognition of inaccessible stored information: Potential bases for incubation effects in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 187–205.

Chapter 14

Metacognition and mindreading: one or two functions?

Joëlle Proust

There is no agreement, in cognitive science or philosophy, about the nature of self-knowledge and its epistemology. No agreement about the functional underpinnings of conscious experience, about the role of emotion in cognition, and about the evolution of the brain. No wonder, then, that at the intersection of all these topics, the function and the scope of metacognition, i.e. cognition about one's cognition, has been hotly debated,[1] and forms the main issue in the present volume. Part of the controversy has to do with the informational processes involved in metacognition. According to a *self-ascriptive view of metacognition* (or SAV), thinkers cannot select, monitor, and control a cognitive activity unless they are also able to reflexively represent *that* they have mental states with specific contents. According to a *self-evaluative view* (or SEV), in contrast, metacognition is one step in a process of active thinking, where agents monitor the available metacognitive feedback in order to adjust their cognitive commands to their cognitive dispositions. In contrast with SAV, however, SEV denies that mindreading is either sufficient or necessary for procedural metacognition. Although procedural metacognition is independent from mindreading, it may be upgraded, when mindreading is available, into analytic or theory-based metacognition.

The first section summarizes the comparative and developmental arguments supporting, respectively, the existence of one or two different functions associated with 'self-knowledge'. Hampton's operational definition of metacognitive behaviour is introduced as an important constraint in the discussion. The second section first examines how, in the abstract, such a function might be fulfilled, through a discussion of a theoretical model, called an *adaptive accumulator module* (AAM). The compatibility of this model with Hampton's definition is discussed, and experimental evidence is presented that AAMs could be major building blocks in procedural metacognition.

In the last section, objections from two angles are addressed. The first claims that procedural metacognition only uses first-order information. The second argues that it engages a form of awareness, which deserves to be classified on a par with analytic, concept-based forms of self-control.

Does metacognition have to involve mindreading?

The case for a single function

The theoretical idea behind SAV is the following. Metacognition, by definition, requires from a creature the capacity to represent cognitive activity, in addition to representing a first-order task

[1] See inter alia: Carruthers (2008, 2009), Proust (2007, 2010).

in which this activity is being exercised (Carruthers 2008). Nelson and Narens's (1990) two-layered schema for metacognition seems prima facie compatible with this definition, although on a theoretical basis rather than as a definition: any good regulator of a system, they insist, must include a model of that system. This architectural constraint, claimed to form a theorem in the mathematics of adaptive control by Conant and Ashby (1970), has long been taken to entail that a second-order representation of the first-order task, at the control level, is a precondition for adequately monitoring a cognitive task.[2] This inference, however, depends on the assumption that a first-order task can only be modelled by a metarepresentation, i.e. a representation attributing to oneself, for example, the belief of being able (with uncertainty *U*) to correctly perform a cognitive task. This assumption has contributed to shaping the intellectualist stance in metacognitive studies, and inspired main-stream research in educational studies.

Another source of inspiration for SAV, however, has come from developmental psychology. Children tested on various forms of cognitive control, self-evaluation, and source monitoring have been shown to have trouble distinguishing the perceptual appearance from the real nature of objects (such as a sponge that looks like a rock) before they reach 4–5 years of age.[3] Similarly with the control and monitoring of memory: children do not seem to try to retrieve events or names before they have understood that they have a mind able to remember. On the basis of such evidence, Josef Perner has persuasively argued that the development of episodic memory in children derives from to the ability to introspect an ongoing experience and interpret it as representing an actual past event.[4] According to him, children do not possess episodic memory until they are able to understand the representational nature of their own minds. As another SAV theorist, Peter Carruthers, puts it, 'It is the same system that underlies our mindreading capacity that gets turned upon ourselves to issue in metacognition'.[5]

The development of epistemic evaluations, furthermore, appears to be more or less parallel with that of mindreading. When 3-year-olds are asked whether they *know* what is inside a box they have never seen before, they, surprisingly, find it difficult to make a reliable judgement. They often answer with a guess, but do not seem to distinguish knowing from guessing before the age of 4 or even later.[6] When asked how long they have known an item of knowledge that was just communicated to them, 3-year-olds regularly respond that they've always known it.[7] In summary, when asked to verbally report about what they know, what appears to them, what they can remember, etc., children seem unable to offer reliable answers before they are able to read their own minds. However, once they have acquired, through verbal communication, the concepts for the basic mental states, and thereby become able to understand how other agents can be wrong about the world, children learn to attribute errors and misrepresentation to themselves as well.[8] It has seemed, then, that cognitive monitoring relies upon the ability to identify one's mental states as such: understanding, first, that people—as well as oneself—have mental states and mental dispositions, that they may or not be correct, and that such correction depends on the amount and quality of evidence available.

---

[2] An alternative interpretation of Conant and Ashby's theorem will be offered at the end of this chapter.

[3] Flavell (1979).

[4] Perner and Ruffman (1995).

[5] See Carruthers (2008, 2009, Chapter 5, this volume). See also Gopnik (1993).

[6] Sodian et al. (2006).

[7] Gopnik and Astington (1988).

[8] Schneider (2008).

The case for two functions

The developmental argument just given, however, has been weakened by three types of findings. First, comparative evidence suggests that non-human primates (including monkeys) present metacognitive competences comparable to those of humans. Granting this result, primate phylogeny should be reflected in human ontogeny, leading us to expect distinctive developmental patterns for self-evaluation and mindreading. Second, recent data indicate that human 3-year-olds indeed present the same metacognitive performances as monkeys, even though they do not yet solve 'false belief task' problems. Third, a series of studies suggest that mindreading is also a biological, rather than a merely cultural ability, which surfaces in various early implicit forms of social sensitivity to others' intentions and beliefs. This hypothesis makes previous correlations between mindreading and metacognition more difficult to interpret, and suggests an independent role for an executive function capacity in both types of performance. We will briefly examine these findings in turn.

Comparative evidence about metacognition as distinct from mindreading

A powerful argument against SAV comes from comparative psychology. Various non-human species that are not adapted to read minds, such as bottlenosed dolphins (*Tursiops truncatus*), and rhesus macaques (*Macaca mulatta*), are able to evaluate whether they are able to discriminate two visual stimuli; they can make a prospective judgement of memory in a serial probe recognition task (Hampton 2001; Smith et al. 2003). In such tasks, the animals are offered the opportunity to 'opt out' from a perceptual or memory task when they feel unable to perform it. The animals' response patterns strikingly resemble those of human subjects. Granting the validity of these experiments, they are compatible with the view that metacognition is a specific adaptation, whose phylogenetic distribution overlaps, but does not coincide, with the ability to read minds. Mindreading is used here to refer to the capacity of identifying beliefs, i.e. what Carruthers and Ritchie (Chapter 5, this volume) call 'stage 2 mindreading', rather than intentions or spatial perspective. Mindreading so understood is a uniquely human ability. Two preliminary issues must be clarified. First, do the methodological difficulties attached to these experiments threaten the validity of this view? Second, supposing that they don't, how could one operationalize the concept of 'metacognition' in non-verbal agents?

Methodological concerns Important methodological concerns have been raised against a hasty metacognitive interpretation of these findings (Carruthers 2008, 2009; Crystal and Foote 2009; Hampton 2009). First, is it not *reward*, rather than the animals' judgements of confidence, that guide decisions? To address this problem, animals were denied any access to reinforcement scheduling, and offered blockwise, rather than trial-by-trial reinforcement. This modification did not affect their metacognitive performance (Smith et al. 2006; Couchman et al. 2010). Second, are not so-called 'metacognitive judgements' actually prompted by *associations between environmental cues*? This worry has been addressed through generalization tests, where the animals need to predict performance in unrelated tasks (Kornell et al. 2007). When the animals immediately transfer their disposition to opt-out, (e.g. from a perceptual to a memory task), it is safe to assume that metacognitive ability is not dependent on the associative strength of the stimuli involved (Hampton 2009; Couchman et al. 2010). Third, are not difficult trials merely *aversive* ones? In a discrimination test, 'middle' stimuli on a continuum might be avoided, not on the basis of a confidence judgement (a judgement of uncertainty), but simply because animals dislike categorizing them (Perner unpublished communication). Several ways of addressing this question have been considered. First, it was shown that capuchin monkeys are able to sort stimuli into three categories, A, B, and middle. However, they are unable to use uncertainty as a motivation to decline

difficult trials on an A–B only task, as rhesus monkeys do, although they thereby incur the cost of long timeouts (Beran et al. 2009). Second, a threshold task, which does not allow a 'middle' category to emerge, has elicited adaptive uncertainty responses in rhesus monkeys (Couchman et al. 2010).

A step forward: an operational definition of metacognition The methodological problems just described have emphasized the need for determining more carefully what counts as a metacognitive capacity. Influenced by SAV, some theorists have claimed that, by definition, metacognition should be based on a secondary *representation*.[9] It is sometimes also claimed that metacognition should be mediated by introspection, with a higher-order conscious state allowing the animal to form a judgement of uncertainty in a trial on the basis of its epistemic feelings. These definitional requirements being difficult to operationalize, however, a less contentious distinction has been offered between a primary and a secondary *behaviour*, or *goal*. Robert Hampton proposes the following list of objective markers for metacognitive behaviour:

1. There must be a primary behaviour that can be scored for its *accuracy*.
2. *Variation* in performance (i.e. uncertainty about outcome) must be present.
3. A secondary behaviour, whose goal is to *regulate* the primary behaviour, must be elicited in the animal.
4. This secondary behaviour must be shown to benefit performance in the primary task (for example, animals must decline tests that they would otherwise have failed).

On top of these four functional conditions, however, Hampton defines a more restrictive metacognitive capacity, which he calls '*private metacognition*' (in contrast with a 'public' form: Hampton 2009). The functional advantage that private metacognition offers is that it enables animals to respond to uncertainty in a generalized way, through endogenous signals, rather than through separately learnt, task-specific associations available to an external observer. The mechanisms for Private metacognition must fulfil three additional, negative conditions.

i. The metacognitive responses must not be based on response competition (where perceptually presented stimuli are merely selected on the basis of their comparative attraction).
ii. They must not be based on environmental cue association.
iii. They must not be based on behavioural cue associations, i.e. 'ancillary responses' such as hesitation, or response latency.

Hampton's three constraints on mechanisms are meant to reveal a capacity for 'private' procedural metacognition. We now have, then, three different candidates for a metacognitive function, that might concurrently fulfil the operational definition: public metacognition (based on publicly available cues), private metacognition (based on internal cues), and mindreading (based on representations of one's mental states). Experimenters aiming to demonstrate private procedural metacognition, Hampton shows, can do so on the basis of a limited number of paradigms. Because it occurs only once a response is given, wagering allows us to disconnect the metacognitive appraisal from the competition of stimuli (condition (i)). By modifying the stimuli involved in the task, transfer tests can control for (ii). Finally, checking on latency times should allow (iii) to be controlled for.

Taking all these conditions together, a few paradigms indeed seem to effectively rule out the effect of exogenous or public influences over metacognitive evaluations. They are the *retrospective gambling paradigm* (also called 'wagering'), and some forms of the *prospective opt-out test*, where

[9] Crystal and Foote (2009b, p. 54).

animals are asked to decide whether or not to perform a task without simultaneously perceiving the test stimuli (Hampton 2009). Animal research thus seems warranted in claiming that private procedural metacognition is manifested in animals that do not have the ability to read their own minds, or other minds.

Developmental evidence favouring a two-function view

Granting that non-humans present procedural metacognition, it would be likely that human children should also do so. Although, as we saw earlier, developmental evidence has long pointed to late development of epistemic self-monitoring—with a schedule parallel to mindreading—it is now realized that the evidence for delayed metacognition might be related to the attributive (or 'explicit') style of most of the tests that were used (Balcomb and Gerken 2008). As we have already seen, children of 3, when tested verbally about what they know (versus what they guess), normally fail to form correct self-attributions of knowledge. However, dissociations frequently occur, in human cognition, between verbal report and behavioural decision. Given the crucial importance of learning and selective information acquisition in our species, it would be very surprising that infants have no sensitivity to the quality of their informational states.[10]

If metacognition is present in young children, as it presumably is in monkeys, a promising method would consist in studying their epistemic behaviour with the paradigms used in comparative psychology. Call and Carpenter (2003), using a set of opaque tubes where food or toys were hidden, showed that 3-year-old children are able to collect information only when ignorant, with performances similar to those of chimpanzees and orangutans. This study, however, did not allow one to determine whether the secondary behaviour was produced by response competition or by access to one's epistemic uncertainty (Hampton 2009). Another option is to use an opt-out paradigm, which is what Balcomb and Gerken (2008) did: they used Smith et al.'s test of memory-monitoring in rhesus monkeys to test children aged 3½. The children first learn a set of paired pictures, representing an animal (target) and a common object (its match). In the subsequent test, they are shown one item of a pair and two possible associates: the match and a distractor; their task is either to select the match, or decline the trial (the stimuli were arranged so that matches and distractors were equally familiar: familiarity could not be used as a cue). Finally, they are given a forced recognition test where they have to select the match of each animal. This study showed that children were adequately monitoring their memory by opting out on the trials they would have failed. A second experiment indicated that they could do so prospectively even when the *only* stimulus presented at the time of decision was the picture of the match (preventing a response competition effect). This experiment thus fulfils the various constraints listed earlier for metacognition. Furthermore, it also seems to offer evidence for 'private metacognition' in children who are not able yet to solve a false belief task.

[10] It is well-known that babies distinguish novel from familiar stimuli: they seem to prefer looking at a familiar object before becoming habituated (before learning), and at a new object thereafter (Hunter et al. 1983). The function of these preferences is clear: adequately targeted cognitive interest allows infants and adults to optimize learning. Another case in point consists in the capacity of 5-month infants to allocate their attentional resources as a function of the type of information they need to extract (for example: species- or property-level information) (Needham and Baillargeon 1993, Xu, 1999). These early types of control of attention, however, do not yet qualify as metacognitive to the extent that the secondary behaviour (appreciating the degree of familiarity with a stimulus) seems to be directly wired into the infant's learning system; as a result, response competition can explain behaviour without invoking a metacognitive decision.

Objection: what if mindreading is a biological, low-level ability?

A series of studies, however, suggesting that mindreading is an early biological, rather than cultural ability, surfacing in various implicit forms of social sensitivity to other's intentions and beliefs, has brought a twist in the one/two-function debate. Onishi and Baillargeon (2005) reported that 15-month-old infants have insight into whether an agent acts on the basis of a false belief about the world. In addition, Kovacs et al. (2010) present evidence that the mere presence of social agents is sufficient, in 7-month-old infants as well as in adults, to automatically trigger online computations about others' goals.[11] As a consequence, mindreading abilities are seen as an innate 'social sense,' that is spontaneous, automatic, and effortless. The relevance of this type of evidence is interpreted differently by SAV and by SEV proponents. SAV proponents, when they take these results as reliable evidence for mindreading,[12] may argue that mindreading, with its early influence on behaviour, is in a position to drive any form of self-evaluation. They need to assume, however, that additional executive and attentional competences explain the late performance of children on high-level, language-dependent tasks such as completing a false-belief task or offering a verbal epistemic self-evaluation.[13] They need, in addition, to downplay the comparative evidence in favour of private metacognition in monkeys.

SEV proponents may argue, in contrast, that if early forms of mindreading are present in infants, then the first appearance, around 4–5 years of age, of metacognitive competences is no longer correlated with, and explainable by, a newly acquired mindreading ability. Delayed metacognition, and delayed false-belief understanding, might be due to extrinsic competences respectively engaged in each function. One way of adjudicating among these two interpretations would involve exploring the mechanisms that might be respectively engaged in metacognition and in mindreading in the human adult.

Do metacognition and mindreading differ in their informational mechanisms?

The most convincing argument in favour of a two-function view would be to show that the informational mechanisms that produce a self-prediction and an other-directed attribution are substantially different, and, to this extent, can produce diverging outcomes. Theorists of noetic judgements have contrasted experience-based and theory-based forms of self-evaluation.[14] Experience consists in feelings, generated by the processes underlying cognitive operations rather than by the agents' attitudes (such as: having a belief) or their outcomes (a belief with a particular content).[15] As we shall see, it can further be hypothesized that the processes that guide self-evaluations in procedural metacognition include a model of the first-order cognitive task; the dynamic properties of the neural vehicle are extracted, and relied upon to model (i.e. monitor and control) the ongoing task. In a nutshell, what makes this model epistemically adequate is that the

[11] This ability belongs to goal prediction, which has been found to be available to infants in their first year (Gergely et al. 1995). Although this ability is sometimes called 'stage-1 mindreading' (Carruthers and Ritchie Chapter 5, this volume), reading a mind is usually defined as a capacity to understand that one's own and others' beliefs can be false.

[12] For an interpretation of Baillargeon's results in terms of behavioural cues, rather than of mindreading, see Perner and Ruffman (2005).

[13] Carruthers (2009).

[14] Koriat and Levy-Sadot (1999).

[15] See Koriat and Levy-Sadot (1999), Schwarz (2002). See Dokic (Chapter 19, this volume) for a discussion of the nature and intentional contents of noetic feelings.

dynamic properties of the vehicle map the epistemic properties of the computational processes involved.

Mindreading-based metacognition, on the other hand, can develop predictions on the basis of a naive theory of the first-order task, and of the competences it engages. The latter thus requires representing both one's own propositional attitudes (such as beliefs and desires) and their contents (that the chocolate is in the drawer). On a two-function view, theoretical metacognition consists in general of knowledge about cognitive dispositions, whereas procedural metacognition is the ability to conduct cue-based self-evaluations. Although mindreading can redescribe and enrich procedural metacognition, it is, from a SEV viewpoint, neither necessary, nor sufficient, to perform contextually flexible metacognitive judgements.

A behavioural dissociation between procedural metacognition and theory-based prediction

According to SAV, the same basic informational processes are involved in self- and other-mental attribution. Therefore knowledge made available to oneself through introspection, or self-directed interpretation, should be automatically transferred to others, and reciprocally: knowledge gained about others should be automatically transferred to self. Results at variance with this prediction have been obtained by Koriat and Ackerman (2010). Participants are asked to memorize—in a self-paced way—pairs of unrelated words. When they have finished learning a given pair, they are asked to offer a judgement of learning (JOL) about their chances to recall this particular pair. This judgement, however, is elicited in two conditions. In condition A–B, the participants first perform the learning task, with a self-evaluative phase after studying each pair (condition A). They then observe another participant performing the task, and are asked to assess the latter's later ability to recall this particular pair (condition B). In condition B–A, the order is reversed: participants first observe another perform the task and predict her success, then perform it themselves.

A simple SAV prediction is wrong on two accounts. First, *the validity of a judgement of learning for a given pair differs* when participants have performed the task before judging, or merely observed another's performance. When they have performed the task, the participants seem to rely on an implicit Memorizing Effort heuristic, that more study time predicts less recall, which turns out to reliably predict successful performance. In contrast, when predicting another agent's ability *before* having performed the task themselves, *subjects rely on a piece of (wrong) folk-theorizing*, that more study time predicts more recall. This suggests that self-evaluation in A elicits a form of procedural, context-sensitive access to the subjective uncertainty associated with a trial, while other-evaluation in B relies on general background conceptual knowledge about successful learning (disregarding the contextual fact that pairs are of unequal difficulty, and that the time spent on a pair reflects that fact).[16]

Second, *transfer turns out to be different* in the A–B and in the B–A conditions. In the A–B condition, the acquisition and transfer to others in B of the metacognitive knowledge acquired in A, in the experimental settings described previously, is found to reliably occur. In the B–A condition, in contrast, participants who, in task B, have merely observed others perform, do not transfer to themselves, in task A, their prediction about others that more time predicts better learning.

[16] There are cases where the dissociation goes the other way round: observers predict more accurately the effects of retention interval for learning in others than in themselves (Koriat et al. 2004). The explanation is the same in both cases, however: procedural metacognition relies on process-based feelings, such as retrieval fluency, which can be a source of illusion, while theory-based control is more prone to involve conceptualizing that time is relevant to prediction of correct retrieval.

The reason they do not, clearly, is that engaging in the metacognitive task themselves allows them to extract additional information that they did not have when merely observing others perform the task.

At this point, SAV theorists might object that a subject, when engaged in a metacognitive task, has access to introspective evidence that she fails to have when she is merely observing another agent. Thus it is expected in SAV terms that (1) the validity of the self-evaluations should differ in the two cases, and (2) that the generalization of knowledge should be asymmetric. In response to this objection, however, note that the participants in the Self condition are unaware of using the implicit effort heuristic. None of them reports, after the experiment, having based their own judgement of learning on an inverse relation between study time and learning. In contrast, participants in the Other condition report having used it to predict learning in others. What does this show? The authors observe that a shift has occurred from experience-based to theory-based JOLs, and that this shift is associated with the need to provide an explicit evaluation of learning in others. Indeed this metacognitive task invites subjects to integrate their own experience with someone else's, which might help the participants to make the underlying effort heuristic explicit. The upshot is that participants do not use the same *kind of knowledge* when predicting learning in others in the A–B and the B–A conditions. In the A–B condition, the knowledge collected in A has its source in the experience generated by a metacognitive engagement. The resulting metacognitive decisions, once made, can subsequently be generalized to another performer based on the subject's general inferential abilities. In the B–A condition, however, the prediction of others' learning relies on a tenet of the naïve theory of memory, according to which longer study time predicts better learning.

Thus a more natural explanation for the dissociation discussed above is that procedural metacognition and mental attribution engage two different types of mechanisms. Engaging in a task with metacognitive demands allows the agent to extract 'activity-dependent' predictive cues, i.e. associative heuristics that are formed as a result of the active, self-critical engagement in a cognitive task. Predicting success in a disengaged way, in contrast, calls forth theoretical beliefs about success in the task. While activity-dependent cues offer a contextual evaluation, theory-laden cues at work in mindreading rely, rather, on conceptual knowledge, which may fail to be sensitive to causally relevant features of potential success in the task.

Additional evidence in favour of this contrast is offered by a third experiment, where the self-other condition is modified. Now participants learning pairs of words in condition A are *not* invited to form a judgement of learning. Will they still apply the memorizing effort heuristic when subsequently predicting learning in others? Interestingly, they fail to do so, with results closely similar to the Other-first condition. This finding, then, suggests that an implicit heuristics is extracted and used only when the task requires making a judgement of learning for each pair. This makes 'activity-dependence' of cue-learning more precise: engagement in self-evaluation, rather than mere engagement in a first-order cognitive task, is a precondition to having the relevant experience, and to transferring it to others.

In summary, an experience of active control-and-monitoring of learning—an idiosyncratic interaction between the learner and the items to be learned associated with an evaluative stance—is needed for subjects to form the correct association between study time and successful retrieval. Transfer to others, however, depends on having conceptually represented the regularity- an ability that might not be available to animals with no such conceptual knowledge. Transfer to others of one's metacognitive experience thus requires mindreading—theorizing about mental states as such—as a necessary step.

The next question, then, concerns the mechanisms that might be selectively engaged in procedural metacognition.

The double accumulator model: theory and evidence

Theory

From classical studies on metacognition and on action, we know that any predictive mechanism needs to involve a *comparator*: without comparing an expected with an observed value, an agent would not be able to monitor and control completion of a cognitive task (Nelson and Narens 1990). When prediction of ability in a trial needs to be made, the agent needs to compare the cues associated with the present task with their expected values. As we saw earlier, these cues can, theoretically, be public. For example, the physical behaviour that is associated with uncertainty (hesitation, oscillation) might be used as a cue for declining a task (which cue, being of a non-introspective kind, is advanced as a reason to favour SAV: see Carruthers 2008).

There are more efficient ways of evaluating one's uncertainty, however, which do not depend on actual behaviour, but only on the informational characteristics of brain activity. The dynamics of activation in certain neural populations can in fact predict—much earlier and more reliably than overt behaviour—how likely it is that a given cognitive decision will be successful. The mechanisms involved in metaperception (i.e. in the control and monitoring of one's perception), described by Vickers and Lee (1998, 2000), have been called *adaptive accumulator modules* (AAMs). An adaptive accumulator is a dynamic comparator, where the values compared are rates of accumulation of evidence relative to a pre-established threshold. The function of this module is to make an evidence-based decision. For example, in a perceptual task where a target might be categorized as an X or as a Y, evidence for the two alternatives is accumulated in parallel, until their difference exceeds a threshold, which triggers the perceptual decision. The crucial information used here consists in the differential rate of accumulation of evidence for the two (or more) possible responses.

Computing this difference—called the balance of evidence—does not yet, however, offer all the information necessary for cognitive control. Cognitive control depends on a secondary type of accumulator, called 'control accumulator'. In this second pair of accumulators, the balance of evidence for a response is assessed against a desired value, itself based on prior levels of confidence associated with that response. Positive and negative discrepancies between the target-level and the actual level of confidence are now accumulated in two independent stores: overconfidence is accumulated in one store, underconfidence in the other. If, for example, a critical amount of overconfidence has been reached, then the threshold of response in the primary accumulator is proportionally reduced. This new differential dynamics provides the system with internal feedback allowing the level of confidence to be assessed and recalibrated over time. [17]

A system equipped to extract this additional type of information can thereby model the first-order task on the basis of the quality of the information obtained for a trial. Genuinely metacognitive control is thus made possible: the control accumulator device allows the system to form, even before a decision is reached, a calibrated judgement of confidence about performance in that trial. Computing the difference between expected and observed confidence helps an agent decide when to stop working on a task (in self-paced conditions), how much to wager on the outcome, once it is reached, and whether to perform the task or not. Granting Vickers and Lee's (2000) assumption that adaptive accumulator modules work in parallel as basic computing elements, or 'cognitive tiles', in cognitive decision and self-evaluation, granting them, furthermore, that the information within each module is commensurable throughout the system, a plausible hypothesis is that these accumulators underlie procedural metacognition in non-humans as well as in humans, in perception as well as, mutatis mutandis, in other areas of cognition.

[17] See Vickers and Lee (1998, p. 181).

Let us check that our four conditions listed earlier are fulfilled by a double-accumulator system. There is clearly a *primary behaviour*, i.e. a primary perceptual or memory task in which a decision needs to be taken. Second, *variation* in performance, i.e. uncertainty in outcome, is an essential feature of these tasks, generated by endogenous noise and variations in the world. Third, the *secondary behaviour*, in control accumulators, consists in monitoring confidence as a function of a level of 'caution': a speed-accuracy compromise for a trial allows the decision threshold to be shifted accordingly. Fourth, secondary behaviour obviously *benefits* performance on the primary task, because it guides task selection, optimizes perceptual intake given the task difficulty, the caution needed and the expected reward, and, finally, reliably guides decision on the basis of the dynamic information that it makes available.

Evidence for adaptive accumulator modules in procedural metacognition

An empirical prediction of AAM models of cognitive control and monitoring bears on how the temporal constraints applying to a task affect a confidence judgement. When the time for which the stimulus is available in a perceptual task is determined by the experimenter—supposing discriminability is constant—the participant's confidence judgement is a direct function of the time for which the stimulus is available (as the prediction is only based on the difference between rates of accumulation for that duration). If, however, the agents can freely determine how long they want to inspect or memorize the stimulus, other things being equal, the prediction is now based on the comparison of the dynamics of the accumulation of the evidence until the criterion is reached, relative to other episodes. Thus, in a self-paced condition, both probability of correctness and associated confidence are *inversely related to the time needed to complete the task* (Vickers and Lee 1998, p. 173). These results are coherent with the research conducted on judgements of learning and judgements of confidence for tasks that have either a fixed, or a self-paced, duration (Koriat et al. 2006).

Further experimental evidence in favour of this theoretical construct comes from the neuroscience of decision-making. Here are a few examples. The first concerns the role of accumulators in metacognitive judgements in rodents. Kepecs et al. (2008) trained rats on a two-choice odour categorization task, where stimuli were a mixture of two pure odorants. By varying the distance of the stimulus to the category boundary, the task is made more or less difficult. Rats were allowed to express their certainty in their behaviour, by opting out from the discrimination task. Conditions 3 and 4 in Hampton's conditions for procedural metacognition are thus met. The neural activity recorded in the orbitofrontal cortex of rats was found to correlate with anticipated difficulty, i.e. with the predicted success in categorizing a stimulus (with some populations firing for a predicted near-chance performance, and others firing for a high confidence outcome). Furthermore, it was shown that this activity did not depend on recent reinforcement history, and could not be explained by reward expectancy. Vickers' control accumulator model offers an explanation: the distance between decision variables, expressed in the differential evolutions in the firing rates, can provide a reliable estimate of confidence in the accuracy of the response. No evidence is collected in this study, however, about the control-accumulator described in Vickers and Lee.[18]

Kiani and Shadlen (2009) also use AAMs to account for the capacity of rhesus monkeys to opt out from a perceptual discrimination task, and choose, instead, a 'sure target' task, on the basis of the anticipated uncertainty of the task. Interestingly, it is activity of populations of neurons in the monkeys' *lateral intraparietal cortex* that was found to represent both the accumulation of

[18] Variance of the decision variables is shown to offer an equivalent basis for confidence judgments, if an appropriate calibration of the criterion value has been made available by prior reinforcement.

evidence, and the degree of uncertainty associated with the decision. The animals, again, satisfy Conditions 3 and 4 in Hampton's list by opting for the sure target when the stimuli were *either* poorly discriminative *or* briefly presented. Moreover, their accuracy was higher when they waived the option than when the option was not available. Finally, a study by Rolls et al. (2010) explores an alternative model for olfactory decisions in humans, 'the integrate-and-fire neuronal attractor network'. This model shares with AAMs the notion that decision confidence is encoded in the decision-making process by a comparative, dynamic cue. Here, the information is carried by differences between increments (on correct trials) and decrements (in error trials) as a function of ΔI (relative ease of decision) of the BOLD signal (i.e. the change in blood flow) in the brain regions involved in choice decision-making. These regions involve, inter alia, the medial prefrontal cortex and the cingulate cortex. This model, however, does not clearly raise the question of how confidence is calibrated, and thus fails to explore the structures allowing metacognitive control.

The models presently used for procedural metacognition tend to suggest, then, that it depends on *two* objective properties of the *vehicle* of the decision mechanisms: first the way the balance of evidence is reached carries dynamic information about the validity of the outcomes; second, the history of past errors, i.e. the observed discrepancies between a target level of confidence and the actual level obtained, carries information about how to adjust the threshold of confidence for a trial, given internal constraints relative to speed and accuracy. Calibration of confidence thus results from a separate dynamic process, storing the variance of the prior positive or negative discrepancies.

In summary, a judgement of confidence is not formed by re-representing the particular content of a decision, or by directly pondering the importance of the outcome. Nor does it require that the particular attitude under scrutiny be conceptually identified (e.g. as a belief). Confidence is directly assessable from the firing properties of the neurons, monitored and stored respectively in the sensory and the control accumulators. A natural suggestion is that metacognitive feelings, such as feelings of perceptual fluency, are associated with ranges of discrepancy in accumulators.[19]

Cognition, procedural metacognition, and mindreading

Proponents of procedural metacognition as well as supporters of a one-function view might reject the present proposal on various, and indeed incompatible, grounds. Some will find the role of AAMs in procedural metacognition compatible with a no-metacognition view, where secondary behaviour is seen as reducible to primary task-monitoring. Others will observe, on the contrary, that adaptive accumulators cannot, as isolated modules, perform all the tasks involved in metacognitive functions. They need to be supplemented by other functional features, such as conscious awareness, attributive and inferential mechanisms, etc., which casts doubt on the claim that procedural metacognition does not need to involve some form of stage-1 mindreading. Finally, it will be observed that the present proposal contrasts two forms of self-knowledge in their respective evolutionary and informational patterns, but does not consider whether, and if so, how, procedural metacognition and mindreading can be integrated into a higher-order form of metacognition.

[19] For lack of space, we will not discuss this suggestion in the present chapter. A theory combining some features of Dokic's 'Water diviner model' and of the 'competence model' (Dokic, Chapter 19, this volume) could explain how feelings generated by accumulator discrepancy can predict likely success in a given cognitive performance.

Objection 1: 'procedural metacognition' boils down to primary task-monitoring

The evidence about AAMs summarized previously might look too close to usual forms of feedback from action to deserve a qualification as metacognitive. If feelings of uncertainty are emergent on the structural properties of decision processes, are they not, finally, 'directed at the world (in particular, at the primary options for action that are open to one), rather than at one's own mental states?', as Carruthers and Ritchie write in this volume (Chapter 5)?[20] From the viewpoint of the animal, it might be that felt uncertainty, or judgements of confidence, are directed at the problem of *how to act in order to get an optimal reward*. In this case, a motivational explanation should be sufficient to account for the kind of monitoring that is supposed to occur in procedural metacognition. A slightly different interpretation of the evidence would claim that the animal feels a conflict between prior expectation and current belief, as in surprise. The existence of such a feeling of conflict, however, does not yet qualify as *metacognitive*. Any emotion, and even any behaviour, will carry information about a primary task; this does not warrant the conclusion that it is metacognitive.[21]

In order to address these objections, it must first be emphasized that the mechanisms assumed to underlie procedural metacognition have an *epistemic* function: this consists in evaluating the validity of a cognitive decision, which contrasts both with a directly *instrumental* function, such as obtaining more reward, and an *executive* function, consisting in allocating more attentional resources to a task. Why might such an epistemic adaptation have evolved? The success of an action—where success is assessed in terms of reward and risk avoidance—presupposes that an organism stores instrumental regularities: in a changing environment, it must be in a position to take advantage of recurring patterns to satisfy its needs. But success of an action also depends on controlling one's cognition, i.e. performing cognitive actions such as directed discriminations or retrievals. This control, however, crucially involves monitoring epistemic deviance with respect to a norm.[22] Just as physical actions are prepared by simulating the act in a context, and need to be evaluated for termination to occur, cognitive actions are prepared by evaluating the probability of the correctness of (and terminated by evaluating the probability of the adequacy of) a given decision. In brief, when predation is high, foraging difficult, or competition high, selective pressure is likely to arise for a capacity to distinguish, on an experiential basis, cases where the world has been changing, or where insufficient information was used to make an epistemic decision. Thus procedural metacognition entails sensitivity to the level of information available; it also entails sensitivity to alternative epistemic norms, such as speed and accuracy, which determine different thresholds of epistemic decision. In contrast with surprise, which is a built-in response meant to increase vigilance, noetic feelings—such as the feeling of confidence—are able to adjust to task and context in a flexible way, as manifested in adequate opting out.

A common mistake in psychophysics consists in failing to distinguish the function of a primary accumulator, which is to make a certainty decision for the current trial, from that of a secondary accumulator, which is to extract the dynamics of error information over successive trials, in order to calibrate the primary accumulator's predictions. The latter function constitutes a different adaptation, as is shown by the fact that, although all animal species have some decision mechanism, few of them monitor the likelihood of error to predictively choose what to do, or to wager about their decision. Indeed the information needed to *make a decision under uncertainty* is not

[20] See also Carruthers (2008).

[21] Carruthers (2008).

[22] See Proust (in press).

the same as the information used in *assessing one's uncertainty*. A decision to do A, rather than B, is made because of A's winning a response competition where the 'balance of evidence' is the basis of comparison. Assessing one's uncertainty, in contrast, relies both on the differential dynamics of the response competition throughout the task, and on an additional comparison between the positive and negative discrepancies between the target and the actual levels of confidence across successive trials. From this analysis about function, we can conclude that an accumulator, potentially, can provide epistemic information, rather than merely carrying it, because it carries it as a consequence of having the function of regulating epistemic decisions: thus the information can be put to use, by a more sophisticated mechanism for controlling epistemic decisions. It appears to be the case that some animals do have such a more sophisticated mechanism.

Now an important question is whether the secondary accumulator, or control accumulator, might be interpreted as metarepresenting the cognitive dispositions manifested in the primary accumulator. Metarepresentation, in general terms, applies to propositional contents attributed under an attitude term to an agent or thinker. Here, no such attributional-propositional process is present. There are, however, interesting similarities and differences between a control-accumulator and a metarepresentation. A metarepresentation offers conceptual information about the content of a mental state, e.g. of a belief; it offers a conceptual model for it. A control-accumulator also models thought; it offers, however, non-conceptual, analogic information about the probability of error/accuracy in confidence judgements, which themselves bear on the outcome of a primary cognitive task. In contrast with metarepresentation, no attitude concept is used in a control accumulator. There is, however, a functional coupling between the primary and secondary accumulators, which guarantees that the secondary accumulator predicts confidence based on evidence in the first, and—through its control architecture—that the second is 'about' the first. This 'aboutness' is reflected in the fact that the noetic feelings are directed at, and concern, the first-order task, i.e. what the animal is trying to do.

Finally, a metarepresentation may allow the organism to predict behaviour, but does not have a fixed rational pattern associated with its predictive potential. Here, in contrast, predictions at the control level immediately issue in adapted cognitive behaviour: information is process-relative, modular and encapsulated. It only allows an agent to adaptively modify its current cognitive behaviour. To explain, and thus remedy persistent discrepancies between expected and observed cognitive success, agents may need to have conceptual knowledge available. Furthermore, various illusions are also created when relying on accumulators to make confidence predictions for abilities they cannot predict (for example, in judging what one will remember at a retention interval on the basis of felt fluency[23]). This narrow specialization of self-prediction is a signature of procedural, as opposed to analytic metacognition.

Objection 2: accumulators are only ingredients in procedural metacognition

A second objection will note, on the contrary, that adaptive accumulators, even if crucial ingredients, are merely ingredients in a larger set of processes involved in metacognition. The indeterminacy of the elements contained in this larger set raises doubts about whether procedural metacognition does not need to involve, for example, stage-1 self-applied mindreading.

It is currently accepted in neuroscience that accumulators are automatic error detection modules, operating in every brain area. Other systems, however, have been proposed to play a role in metacognitive regulation and control. A 'conflict monitoring system', located in the anterior

[23] Cf. Koriat et al. (2004).

cingulate cortex, is known to have the function of anticipating error and correcting it on line. This system is based not on confidence judgements and control accumulators, but on the fact that working memory can activate processing pathways that interfere with each other (by using the same resources or the same structures), a situation which makes processing unreliable.[24] Furthermore, an analytic, conscious, deliberate conceptual system has been found, in humans, to contribute to metacognitive judgement, and sometimes to override confidence judgements resulting from the procedural metacognition.[25] This documented variety of mechanisms, however, does not warrant the one-function view. Rather, it emphasizes the phylogenetic difference between procedural and analytic metacognition. The first type relies on a variety of mechanisms to detect and control error; the second is a distinct adapation, which enables agents to understand error as false belief.

The neurophysiological and experimental evidence discussed above, furthermore, suggests that feelings of confidence are not mediated by a conception of the self nor by higher-order attributional mechanisms. In accord with this evidence, it should be stressed that the brain areas respectively involved in metacognition and in mindreading do not seem to overlap.[26] The first include, in humans, the sensory areas (primary accumulators), the dorsolateral prefrontal cortex and the ventro-medial prefontal cortex, in particular area 10 (where control accumulators may be located) and the anterior cingulate cortex. Lesion studies show that the right medial prefrontal cortex plays a role in accurate feeling-of-knowing judgements.[27] Transcranial magnetic stimulation applied to the prefrontal cortex has been further shown to impair metacognitive visual awareness.[28] Mindreading, in contrast, involves the right temporal-parietal junction, the prefrontal antero-medial cortex, and anterior temporal cortex.[29]

Another argument can be drawn from a behavioural phenomenon called 'immunity to revision of noetic feelings'. In a situation where subjects become aware that a feeling has being produced by a biasing factor, they are in a position to form an intuitive theory that makes subjective experience undiagnostic. In such cases, the biased feelings can be controlled for their effect on decision.[30]. The experience itself, however, survives the correction.[31] Why does experience present this strange property of immunity to correction in the face of evidence?

Nussinson and Koriat (2008) speculate that noetic feelings involve two kinds of 'inferences'.[32] In a first stage, a 'global feeling', such as a feeling of fluency, is generated by 'rudimentary cues concerning the target stimulus', which are activity-dependent.[33] In a second stage, a new set of cues are now identified in the light of available knowledge about the stimulus, the context, or the

[24] Botvinick et al. (2001).

[25] Koriat and Levy-Sadot (1999).

[26] I am deeply indebted on this matter to Stan Dehaene's Lectures on metacognition at the Collège de France, Winter 2011.

[27] Schnyer et al. (2004), Del Cul et al. (2009).

[28] Rounis et al. (2010).

[29] Perner and Aichorn (2008).

[30] Unkelbach (2007) shows, for example, that participants can attribute to the same feeling of fluency a different predictive validity in a judgment of truth.

[31] Nussinson and Koriat (2008).

[32] It may be found misleading to use the same term of 'inference' for an unconscious predictive process, which seem to rely on the neural dynamics of the activity or, as the authors hypothesize, on implicit heuristics, and for a conscious, conceptual process, which can integrate the subject's knowledge about the world.

[33] In the interpretation offered here, the implicit cues and heuristics ultimately consist in the dynamics of the paired accumulators.

operation of the mind. A new judgement occurs using conscious information to interpret experience. The imperviousness of experience to correction might thus be causally derived from the automatic, unconscious character of the phase 1 processing that generates it. Such a two-stage organization of feelings, and the fact that the experience and associated motivation to act cannot be fully suppressed or controlled, speak in favour of our two-function view.

Conclusion

This chapter has defended a two-function view of self-knowledge. One function consists in procedural metacognition, a capacity that has been proposed to depend crucially on the coupling of control and monitoring accumulator mechanisms. Blind to contents, this form of self-evaluation takes as its input dynamic features of the neural vehicle, and yields practical epistemic predictions as output, concerning whether the system can, or cannot, meet a normative standard in a given cognitive task. It is, thus, contextually sensitive to attitudes and to their associated conditions of correction. The other source of self-knowledge is conceptual; mindreading offers human beings a conceptual understanding of their own cognitive dispositions, which in turn allows them to override, when necessary, the decisions of procedural metacognition. These two routes to self-knowledge have a parallel in the so-called 'dual-process theory' of reasoning, where 'System 1' is constituted by quick, associative, automatic, parallel, effortless, and largely unconscious heuristics (such as the availability heuristics), while 'System 2' encompasses slow, analytic, controlled, sequential, effortful, and mainly conscious processes.[34] The present discussion suggests that self-evaluation might similarly depend on two such systems. Noetic feelings seem to be the subjective, emotional correlates of subpersonal accumulator features such as neural latency, intensity and stability; they are also immune to revision: all features associated with System 1. If they deliver consistently inappropriate predictions, i.e. produce metacognitive illusions, controlled processing of System 2 is supposed to step in, as its presumed function is to 'decontextualize and depersonalize problems'.[35] An open question remains, at this point: is such stepping-in entirely dependent on a mindreading capacity? The present state of the literature suggests a positive answer, but comparative psychology might still surprise us.

Acknowledgements

I am grateful to Richard Carter both for his linguistic help and his comments, and to Michael Beran, David Smith, and Kirk Michaelian, for their critical observations on a prior draft of this chapter.

References

Balcomb, F. K. and Gerken, L. (2008). Three-year old children can access their own memory to guide responses on a visual matching task. *Developmental Science*, 11(5), 750–60.

Beran, M. J., Smith, J. D., Coutinho, M. V. C., Couchman, J. J., and Boomer, J. (2009). The psychological organization of 'uncertainty' responses and 'middle' responses: A dissociation in capuchin monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 371–81.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3), 624–52.

[34] Stanovitch and West (2000).

[35] Stanovitch and West (2000, p. 659).

Call, J. and Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition*, 4, 207–20.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind and Language*, 23, 58–89.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32, 121–38.

Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1, 89–97.

Couchman, J. J., Coutinho, M. V. C., Beran, M. J., and Smith, J. D. (2010). Beyond stimulus cues and reinforcement signals: A new approach to animal metacognition. *Journal of Comparative Psychology*, 124(4), 356–68.

Crystal, J. D. and Foote, A. L. (2009a). Metacognition in animals. *Comparative Cognition and Behavior Reviews*, 4, 1–16.

Crystal, J. D. and Foote, A. L. (2009b). Metacognition in animals: Trends and challenges. *Comparative Cognition and Behavior Reviews*, 4, 54–5.

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., and Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, 132(9), 2531–40.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* 34(1979), 906–11.

Flavell, J. H., Green, F. L., and Flavell, E. R. (1995). Young children's knowledge about thinking. *Monographs of the Society for Research in Child Development*, 60(1, Serial No. 243).

Gergely, G., Nadasdy, Z., Csibra, G. and Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition* 56, 165–93.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16(1), 1–14, 29–113.

Gopnik, A. and Astington, J. W. (1988). Children's understanding of representational change and it relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, 59(1), 26–37.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 5359–62.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comparative Cognition and Behavior Reviews*, 4, 17–28.

Hare, B., Call, J., Agnetta, B., and Tomasello, M., (2000). Chimpanzees know what conspecifics do and do not see, *Animal Behaviour*, 59, 771–85.

Hunter, M., Ames, E., and Koopman, R. (1983). Effects of stimulus complexity and familiarization time on infant preferences for novel and familiar stimuli. *Developmental Psychology*, 19(3), 338–52.

Kepecs, A., Naoshige, U., Zariwata, H., and Mainen, Z.F. (2008). Neural Correlates, computation and behavioural impact of decision confidence. *Nature*, 455, 227–31.

Kiani, R. and Shadlen, M. N. (2009). Representation of Confidence associated with a decision by neurons in the parietal cortex. *Science*, 324(5928), 759–64.

Koenig, M. A. and Echols, C. H. (2003). Infant's understanding of false labeling events: the referential roles of words and the speakers who use them. *Cognition*, 87, 179–208.

Koriat, A. (2000). The feeling of knowing: some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149–71.

Koriat, A. and Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for Self and Other during self-paced study. *Consciousness and Cognition*, 19(1), 251–64.

Koriat, A. and Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken and Y. Trope (Eds.) *Dual Process Theories in Social Psychology*, pp. 483–502. New-York: Guilford.

Koriat, A., Bjork, R. A., Sheffer, L., and Bar, S. K. (2004). Predicting one's forgetting: the role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, 133(4), 643–56.

Koriat, A., Ma'ayan, H., and Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General:* 135(1), 36–69.

Kornell, N., Son, L., and Terrace, H. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18, 64–71.

Kovács, A. M., Téglás, E., and Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012), 1830–34.

Loussouarn, A., Gabriel, D., and Proust, J. (2011). Exploring the informational sources of metaperception: The case of Change Blindness Blindness. *Consciousness and Cognition*, 20,1489–1501.

Marazita, J. M. and Merriman, W. E. (2004). Young children's judgment of whether they know names for objects: the metalinguistic ability it reflects and the processes it involves. *Journal of Memory and Language*, 51, 458–72

Needham, A. and Baillargeon, R. (1993). Intuitions about support in 4.5-month-old infants. *Cognition*, 47, 121–48.

Nelson, T. O. and Narens, L. (1990). Metamemory: a theoretical framework and new findings. In T. O. Nelson (Ed.) (1992) *Metacognition, Core Readings*, pp. 117–30. Boston, MA: Allyn and Bacon.

Nussinson, R. and Koriat, A. (2008). Correcting experience-based judgments : the perseverance of subjective experience in the face of the correction of judgment. *Metacognition Learning*, 3, 159–74.

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308, 255–8.

Perner, J. and Aichorn, M. (2008). Theory of mind, language and the temporo-parietal junction mystery. *Trends in Cognitive Sciences*, 12(4), 123–6.

Perner, J. and Ruffman, T. (1995). Episodic memory and autonoetic consciousness: Developmental evidence and a theory of childhood amnesia. Special Issue: Early memory. *Journal of Experimental Child Psychology* 59(3), 516–48.

Perner, J., Kloo, D., and Stöttinger, E. (2007). Introspection and remembering. *Synthese*, 159, 253–70.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Perner, J. and Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences*, 3(9), 337–44.

Perner, J. and Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, 308, 214–16.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition?' *Synthese*, 2, 271–95.

Proust, J. (2010) Metacognition. *Philosophy Compass*, 5(11), 989–98.

Proust, J. (in press). Mental acts as natural kinds. In A. Clark, J. Kiverstein, and T. Vierkant (Eds.) *Decomposing the Will*. Oxford: Oxford University Press.

Rolls, E. T., Grabenhorst, F., and Deco, G. (2010). Decision-making, errors, and confidence in the brain. *Journal of Neurophysiology*, 104, 2359–74.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., and Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, 1(3), 165–75.

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain and Education*, 2, 114–21.

Schneider, W. and Lockl, K. (2002). The development of metacognitive knowledge in children and adolescents. In T. J. Perfect and B. Schwartz (Eds.) *Applied Metacognition*, pp. 224–57. Cambridge: Cambridge University Press.

Schnyer, D. M., Verfaellie, M., Alexander, M. P., LaFleche, G., Nicholls, L. and Kaszniak, A. W. (2004). A role for right medial prefrontal cortex in accurate feeling-of-knowing judgments: evidence from patients with lesions to frontal cortex. *Neuropsychologia*, 42, 957–66.

Schwarz, N. (2002). Situated cognition and the wisdom of feelings: Cognitive tuning. In L. F. Barrett and P. Salovey (Eds.) *The wisdom in feelings: Psychological processes in emotional intelligence*, pp. 144–66. New York: Guilford.

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R. and Erb, L. (1995). The uncertain response in the bottlenosed dolphin *Tursiops truncates*. *Journal of Experimental Psychology: General*, 124, 391–408.

Smith, J. D., Shields, W. E., and Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26, 317–73.

Smith, J. D., Beran, M. J., Redford, J. S., and Washburn, D. A. (2006). Dissociating uncertainty states and reinforcement signals in the comparative study of metacognition. *Journal of Experimental Psychology: General*, 135, 282–97.

Smith, J. D., Beran, M. J., Couchman, J. J., Coutinho, M. V. C., and Boomer, J. B. (2009). Animal metacognition: Problems and prospects. *Comparative Cognition & Behavior Reviews*, 4, 40–53.

Sodian, B., Thoermer, C., and Dietrich, N. (2006). Two- to four-year old children differentiation of knowing and guessing in a non-verbal task. *European Journal of Developmental Psychology*, 3: 222–37.

Sodian, B. and Wimmer, H. (1987). Children's understanding of inference as a source of knowledge. *Child Development*, 58, 424–33.

Stanovich, K. E. and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645–65.

Surian, L. and Leslie, A. M. (1999). Competence, performance in false belief understanding: a comparison of autistic and three year-old children. *British Journal of Developmental Psychology*, 17, 141–55.

Unkelbach, C. (2007). Reversing the Truth Effect: learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 33(1), 219–30.

Vickers, D. and Lee, M. D. (1998). Dynamic Models of Simple Judgments: I. Properties of a Self-Regulating Accumulator Module. *Nonlinear Dynamics, Psychology and Life Sciences*, 2(3), 169–94.

Vickers, D. and Lee, M. D. (2000). Dynamic Models of Simple Judgments: II. Properties of a Self-Organizing PAGAN Model for multi-choice tasks. *Nonlinear Dynamics, Psychology and Life Sciences*, 4(1), 1–31.

Washburn, D. A., Smith, J., and Shields, W.E. (2006). Rhesus monkeys (*Macaca mulatta*) immediately generalize the uncertain response. *Journal of Experimental Psychology: Animal Behavior Processes*, 32, 85–9.

Xu, F. (1999). Object individuation and object identity in infancy: the role of spatiotemporal information, object property information, and language. *Acta Psychological; Special Issue: Visual object perception*, 102(2–3), 113–36.

Chapter 15

Metacognition and indicative conditionals: a précis

Hannes Leitgeb

In this paper we aim to defend the following claim:

- Thesis: *accepting an indicative conditional is a metacognitive process that is not metarepresentational.*

In other words: first of all, the mental process of accepting an indicative conditional is not an instance of cognition about the external world. It is an instance of cognition *about cognition*, but, as we are going to argue, one that is *not* metarepresentational either, as it does not involve a representation of that process *as being mental*. So the acceptance of an indicative conditional is metacognitive without being metarepresentational, that is, it is metacognitive in the specific sense of Proust (2007). We will have to restrict ourselves to a sketch of the argument in favour of this thesis—the fully worked out version of the argument will have to be left for a different paper.

In order to clarify and defend the thesis, we will need to turn in more detail to the two central locutions that are contained in it: 'accepting an indicative conditional' and 'metacognition'. The first section in this chapter ('Accepting an indicative conditional') explicates how a rational agent plausibly goes about to accept an indicative conditional and how this kind of acceptance can be made precise in formal terms. From this it will follow in the second section ('The impossibility result and its consequences') that accepting an indicative conditional differs fundamentally from believing a proposition to be true, whether the belief is one about the world or an introspective one. The third section ('Metacognition and the defence of the thesis') then explains what a metacognitive state is by building on work done by Proust (2007). In the course of that explanation, we will also develop our argument for the thesis from above, based on our previous considerations. The final section ('Open questions') concludes the paper with some open questions that will have to be left to a more comprehensive study of metacognition and the acceptance of conditionals.

Accepting an indicative conditional

It is well known that the subjective acceptability of a conditional 'If *A* then *B*' does not merely depend on the propositional contents of its antecedent and its consequent, but also on the grammatical mood in which its antecedent and consequent are formulated. For instance, using Ernest Adams' famous example, the indicative conditional:

- If Oswald didn't kill Kennedy, someone else did.

is acceptable to any person informed of the Kennedy assassination, while this is not so for the corresponding subjunctive conditional (counterfactual):

- If Oswald hadn't killed Kennedy, someone else would have.

In the following, we will concentrate solely on indicative conditionals: these are typically formulated in natural language by means of 'did-did' or 'does-will' constructions, and in contrast with subjunctive conditionals, they are such that if asserted, the act or state that is described by the antecedent is assumed to be actually the case—'If Oswald [*actually*] didn't kill Kennedy, then…'—rather than just being contemplated as a mere possibility.

This 'assumed to be actually the case' character of the antecedents of indicative conditionals is also the starting point of what is maybe *the* leading philosophical theory of indicative conditionals these days: the so-called *Suppositional Theory of indicative conditionals*, which has been defended—in different variants, and amongst others—by Ernest Adams, Dorothy Edgington, Vann McGee, Jonathan Bennett, and Isaac Levi. (A nice overview of the theory can be found in Bennett 2003, chapters 4–9.) The guiding thought behind the suppositional theory is the famous Ramsey test for conditionals, which goes back to a brief remark made by Frank P. Ramsey (1929 p. 155 fn) in his 'General propositions and causality':

> If two people are arguing 'If *p* will *q*?' and are both in doubt as to *p*, they are adding *p* hypothetically to their stock of knowledge and arguing on that basis about *q*.

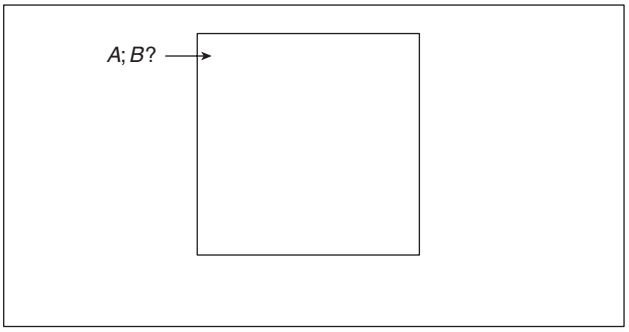
The same content is conveyed by R. Stalnaker's (1968, p. 102) slightly more detailed reformulation:

> This is how to evaluate a conditional: First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.

There are four main aspects to this Ramsey test procedure:

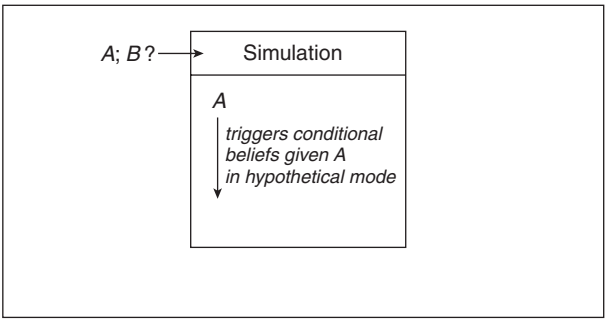
1. The acceptance of an indicative conditional is tied to a form of *suppositional reasoning* that takes the antecedent of the conditional as its input (cf. 'adding p', 'add the antecedent' in the earlier quotes).
2. The means and resources by which this type of suppositional reasoning is being realized are the very same ones that are used when an agent draws an *inference* or *revises* her beliefs on the basis of actual evidence (cf. 'arguing on that basis', 'adjustments', 'maintain consistency'). That is also why the supposition of the antecedent is really a *supposition-as-a-matter-of-fact*, and why 'assuming the antecedent to be actually the case' is much like learning actual evidence.
3. However, unlike proper inferences or revisions of belief that are based on evidence, the suppositional reasoning process in question applies these means and resources *off-line*, in terms of a kind of a simulation (cf. 'hypothetically'). Therefore, when the suppositional reasoning process leads ultimately to a positive appraisal of the consequent—on the supposition of the antecedent—this does not cause the agent to actually believe the consequent outside of the simulation context.
4. The Ramsey test process that initiates the suppositional reasoning process, and which assesses its outcomes, is *external* to the means and resources that are involved in (2) and (3) (cf. 'arguing "If p will q?"', 'evaluate a conditional'). If in the simulation the consequent is accepted on the supposition of the antecedent, then *outside* of the simulation it is the *conditional* that is accepted *simpliciter*.

This is how the Ramsey test unfolds in terms of stages. Assume that a Ramsey test procedure *X* for an indicative conditional $A \rightarrow B$ is initiated within an agent's cognitive system:



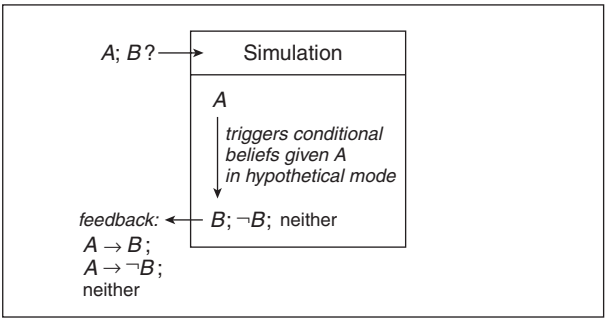
In the picture, the Ramsey test *X* is located outside of the inner square, while the suppositional reasoning process *Y* that will be triggered by it will be represented by what is going on in the inner square. *X* is has both *A* and *B* available; its task is to determine the acceptability of $A \rightarrow B$ on the basis of the process *Y* to which *A* will be handed as its input and which is called upon in order to deliver an assessment of *B* on the basis of that input. *Y* is a suppositional reasoning process from the viewpoint *external* to the inner square: for its input *A* is merely supposed to be the case, and *Y*'s ultimate verdict on *B* will not be translated ultimately into an attitude towards *B*; but if looked at from *within* the inner square, *Y* is going to behave more or less just like a normal inference or revision process.

Next, *X* starts that process *Y* in simulation mode: *Y* is fed with *A as if* the agent had acquired the actual belief in *A*, and within the inner square the role of *A* is indeed more or less the same as the role of an actual belief content—all relevant beliefs the agent has conditional on *A* become activated, more or less just as it would be the case if *A* were a new piece of evidence that would trigger various relevant inferential processes which would deliver new beliefs on the basis of that evidence. (This is a simplification that is not quite justified for *introspective* statements *B* as in the famous Thomason conditionals, such as 'If my wife is cheating in me, then I don't know it'. But this is not much of a worry for most consequents *B* that are about the external world. See Leitgeb (2011) for a detailed treatment of the problematic introspective cases; and for more information on conditional beliefs in general, i.e. on *beliefs conditional on a proposition*, see Leitgeb (2007).) But since from *X*'s viewpoint, that is, outside of the inner square, this is all part of a simulation, this 'belief' in *A* is nevertheless a merely hypothetical one:



As soon as *Y* has done its job of assessing *B* on the basis of *A*—which may lead, if stated in categorical terms, to a positive appraisal of *B*, or a negative appraisal of *B* (that is, a positive appraisal

of $\neg B$), or to neither of them—the result is being fed back to X which translates it into a corresponding appraisal of the original conditional: either $A \rightarrow B$ is accepted, or $A \rightarrow B$ is rejected (that is, $A \rightarrow \neg B$ is accepted), or neither of the two, respectively, where the last of these outcomes would mean that the agent is indifferent with respect both to the acceptability of $A \rightarrow B$ and $A \rightarrow \neg B$:



This ends the Ramsey test procedure.

Note that the Ramsey test in this form seems to be valid only for indicative conditionals: the supposition that Oswald did not kill Kennedy leads to a revised state in which it is indeed believed (hypothetically) that someone else must have killed Kennedy; accordingly, 'If Oswald didn't kill Kennedy, someone else did' is accepted. But it is not possible to explain *in the very same way* the *lack* of acceptance of the subjunctive 'If Oswald hadn't killed Kennedy, someone else would have'. Supposing as a matter of fact that Oswald did not kill Kennedy cannot just by itself do the trick, or otherwise the indicative conditional would not have been acceptable in the first place. In order for a version of the Ramsey test to hold also for conditionals such as 'If Oswald hadn't killed Kennedy, someone else would have', the manner in which supposition is being implemented would have to be changed to something like subjunctive or counter-to-the-fact supposition (see section 2 of Leitgeb 2012). But since we will only deal with indicative conditionals in the following, this need not concern us here.

How can the Ramsey test for indicative conditionals be stated in more precise formal terms? First of all, an appropriate scale of measurement needs to be chosen that determines whether the concept of the acceptability of an indicative conditional is actually a categorical 'all-or-nothing' concept, or whether it is a concept on an ordinal or even on a numerical scale. For instance, Gärdenfors (1988) suggests a qualitative formalization in terms of a so-called belief revision operator * which operates on sentences (the input A) and deductively closed belief sets K of sentences (the agent's present state of belief), and which determines whether after revising K in the light of A the sentence B ends up being believed, that is, whether $B \in K^\star A$. If this is so, and only if it is, $A \rightarrow B$ is being accepted. The corresponding revision operator * is assumed to satisfy plausible rationality postulates (see Gärdenfors (1988) for the details), which can be stated in precise logical terms, and which by this version of the Ramsey test translate into plausible postulates on the acceptability of conditionals (cf. Leitgeb 2010). The resulting theory thereby yields a formalization of the Ramsey test on a categorical scale.

But the best-known formalization of the Ramsey test is actually a quantitative one which is spelled out in terms of the *subjective probability* of sentences: Let $P(A)$ be the degree of belief in the sentence A—a number in the real number interval $[0,1]$—as being given by a degree-of-belief function P which measures the strength with which a particular agent believes in the truth of the sentences of some language at some given point of time. Assuming this function P is a probability

measure means that P is taken to satisfy the axioms of the probability calculus: for instance, by the laws of probability, $P(A \vee \neg A) = P(A) + P(\neg A) = 1$; that is: the degree of belief in the tautology $A \vee \neg A$ is 1, which is the maximal possible degree of belief, and the degree of belief in the negation of A is nothing else than 1 minus the degree of belief for A. There are strong arguments for the thesis that the degree-of-belief function of any *rational* agent must obey the laws of probability (e.g. the so-called Dutch Book arguments), but we will not discuss this in any further detail. Instead, let us simply take for granted that the agents that we will be dealing with do have numerical degrees of belief which they distribute over sentences in line with a probability measure. Such a probability or degree of belief $P(A)$ in the sentence A is then also called the absolute or unconditional probability of A (as being given by P).

Once such an absolute or unconditional probability measure is in place, it is also possible to define *conditional* probabilities in terms of it: formally, the conditional probability of B given A—that is, the probability of B *on the supposition of* A—as being determined by the absolute probability measure P is denoted by:

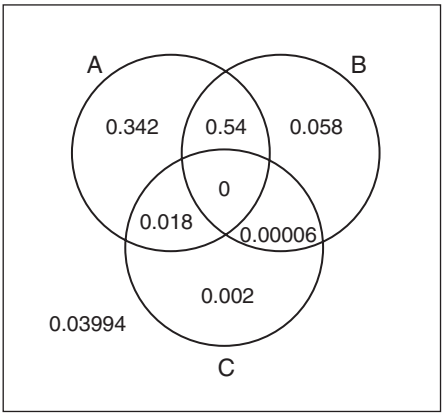
$$P(B \mid A)$$

And it is formally defined by means of a ratio of probabilities:

$$P(B \mid A) = P(A \wedge B) \,/\, P(A)$$

(Of course, this ratio is only well-defined if the probability of A is not equal to 0.)

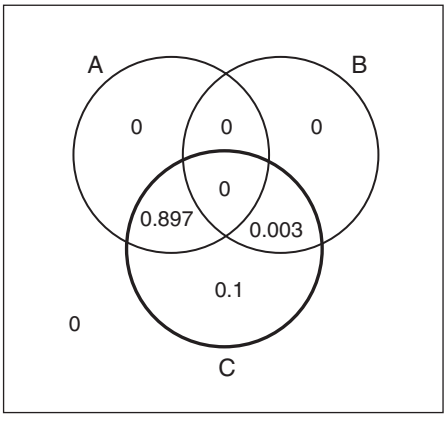
In terms of a simple example: let us assume a probability measure distributes degrees of belief over all possible Boolean combinations of three sentences A, B, C as depicted by the following Euler-Venn diagram:



Hence, for instance, the probability of A is 0.342 + 0.018 + 0.54 + 0 = 0.9, the probability of not-A is then 0.1 of course, the probability of B is 0.54 + 0.058 + 0.00006 + 0 = 0.549806, the probability of A & B is 0.54 + 0 = 0.54, and so forth. Now let us assume that we want to determine the conditional probability of some sentence *given* C: then what the ratio formula amounts to is that it makes us zoom onto the area that corresponds to C pretending that this area is the complete space of possibilities; accordingly, all probabilities of sentences that contradict C—the probabilities of all areas that lie outside of the area of C—are set to 0, and the probabilities of sentences that are consistent with C—the probabilities of all areas that lie inside of the area of C—are multiplied with one and the same constant, so that overall this new distribution of degrees of belief

conditional on C satisfy the axioms of the probability calculus again. Any such process that leads from a probability measure such as *P* and a sentence such as *C* to a new probability measure that is defined in terms of conditional probabilities given *C* is called *conditionalization* (of *P* on *C*).

In our example, conditionalization of *P* on *C* yields:



In a straightforward sense, this procedure is but the quantitative analogue of what we described to be going on in the Ramsey test procedure *inside of the inner square*: in Stalnakerian terms, the sentence *C* is first added to one's stock of beliefs, that is, in the present context, the agent's present probability space gets restricted to the area of *C*. That is: the actual world is supposed to lie somewhere in that area. Secondly, the necessary adjustments are being made that are required to maintain probabilistic coherence so that once again a probability measure emerges from this. Finally, the resulting degree of belief of a sentence can be determined. For instance, in the example above, the degree of belief in *B* on the supposition of *C* that is determined in this way is $0.003 + 0 = 0.003$. In this case, the supposition of *C* leads to a drastic decrease of the probability of *B* from 0.549806 originally to 0.003 after conditionalization. On the other hand, the probability of $\neg B$ is bumped up to 0.997 by conditionalization on *C*. And just as intended by the Ramsey test, the very procedure that is applied inside the inner box would also be applied to actual evidence if it were to be incorporated in an agent's belief system: indeed, probabilistic conditionalization is not just used as a formalization of *supposition*, it is also *the* standard Bayesian method of *update* or *learning* in the light of evidence.

It was E. Adams (1975) who first suggested that conditionalization may also be applied in order to give a probabilistic reconstruction of the Ramsey test by which numerical degrees of acceptability for indicative *conditionals* are being determined:

- Probabilistic Ramsey Test (Adams' thesis):

  For every subjective probability measure *P* that an agent can possibly have, for all sentences *A*, *B* (with $P(A) > 0$, and where we will assume *A* and *B* are restricted so that they do not themselves contain the indicative conditional sign $\rightarrow$):

  The degree of acceptability of $A \rightarrow B$ (in *P*) equals $P(B \mid A)$.

For instance, in our example from before: the degree of acceptability of the indicative conditional $C \rightarrow B$ is 0.003, while the degree of acceptability of $C \rightarrow \neg B$ is 0.997. So, the agent would

pretty much accept $C \rightarrow \neg B$ and more or less reject $C \rightarrow B$, by the probabilistic Ramsey test. In this manner, the original Ramsey test can be explicated (in Rudolf Carnap's sense of explication) in precise numerical terms by incorporating it into a well-defined system of scientific concepts, that is, in this case, probability theory. But this comes with a surprising consequence, as the next section is going to show.

The impossibility result and its consequences

It might seem that one could be more specific in our probabilistic reconstruction of the Ramsey test than we were in the earlier formulation: for what else than the *degree of belief in the sentence* $A \rightarrow B$ *being true* could be meant by the 'degree of acceptability of $A \rightarrow B$'? In other words: We should be able to denote the degree of acceptability of $A \rightarrow B$ (in P) by '$P(A \rightarrow B)$', where P would be applied to the sentence $A \rightarrow B$ in precisely the same sense, and subject to the same formal constraints—the axioms of absolute or unconditional probability—as this is the case in '$P(A)$', '$P(B)$', '$P(A \,\&\, B)$', and the like.

This proposal, which was put forward first by R. Stalnaker (1970) and which certainly looks plausible prima facie, leads to yet another probabilistic reconstruction of the Ramsey test:

- Probabilistic Ramsey Test II (Stalnaker's thesis):

  For every subjective probability measure P that an agent can have, for all sentences A, B (with $P(A) > 0$ and the same restrictions on A and B as before):

  $$P(A \rightarrow B) = P(B \mid A)$$

  where $P(A \rightarrow B)$ is the absolute or unconditional probability of $A \rightarrow B$.

However, as it turns out, in spite of its prima facie plausibility, Stalnaker's thesis *cannot* be true:

> Theorem: The Probabilistic Ramsey Test II (Stalnaker's Thesis) for conditionals is inconsistent with (i) the standard axioms of probability, together with (ii) non-triviality assumptions on the possible subjective probability measures of an agent. (Lewis 1976.)

As David Lewis was able to prove, Stalnaker's thesis can only be satisfied by all the probability measures that an agent could possibly have if that class of probability measures is trivialized, e.g. if the probability measures in that class only take 0s and 1s as their values or are similarly severely restricted in their range of values. And if triviality in this sense is excluded by assumption, then this assumption together with the standard axioms of probability and with Stalnaker's thesis leads to plain inconsistency, which is precisely what the earlier theorem expresses.

What is the proper interpretation of this surprising result? As always, there is more than just one way to go here, but probably the most obvious conclusion is: assume that the original Probabilistic Ramsey Test (Adams' thesis) from before is correct. The only additional assumption that was needed in order to move from it to the flawed Probabilistic Ramsey Test II (Stalnaker's thesis) was that the degree of acceptability of $A \rightarrow B$ is identical to the degree of belief in $A \rightarrow B$ being true. So this latter assumption must be false. Now, in line with the Probabilistic Ramsey Test, $A \rightarrow B$ certainly does have a degree of acceptability as being given by its corresponding conditional probability. It is furthermore plausible to assume that if $A \rightarrow B$ also had a subjective probability in the normal sense of the word—a probability of being true—then this probability should be equal to the degree of acceptability of $A \rightarrow B$. But that cannot be the case by Lewis' theorem. Hence: *$A \rightarrow B$ does not even have a probability of being true.*

In other words: *indicative conditionals do not express propositions, for otherwise accepting them should coincide with believing them (to be true). Therefore, indicative conditionals cannot have truth*

values either, although they do come with subjective values of acceptance as given by the Probabilistic Ramsey Test (Adams' thesis). And indeed this is exactly what the Suppositional Theory of indicative conditionals holds.

We should add that these considerations are robust under altering the scale level of acceptability: even if we had opted for Gärdenfors' qualitative reconstruction of the Ramsey test by means of a so-called belief revision operator, an impossibility result that is quite similar to Lewis' could have been derived from the conjunction of Gärdenfors' Ramsey test, a non-triviality assumption on an agent's possible belief revision operators, and the assumption that the acceptance of a sentence coincides with the belief in the truth of the sentence: as was proven by Gärdenfors (1986). A similar diagnosis as before would have led to the same conclusion that indicative conditionals do not express propositions, for otherwise the acceptance of an indicative conditional should be a belief in a proposition which cannot be the case by Gärdenfors' theorem.

This leaves us with the following obvious question: if an indicative conditional $A \rightarrow B$ does not express a proposition, *then what is the communicative purpose of an agent's assertion of* $A \rightarrow B$? It cannot be that an assertion of $A \rightarrow B$ says or reports that the world is such-and-such; for then $A \rightarrow B$ would have to express a proposition, which was excluded by our previous considerations. Furthermore: let '*S*' be a proper name of the agent who asserts $A \rightarrow B$: then it cannot be the case either that *S*'s assertion of $A \rightarrow B$ says or reports that *S* is in such-and-such a state of mind; for example, that *S*'s conditional degree of belief in *B* given *A* is high, or the like. For if so, then once again $A \rightarrow B$ would have to express a proposition—in this case, a proposition about *S*'s mental state—which cannot be the case by the earlier given theorems.

There is a remaining 'quasi-propositional' option: when we said before that Lewis (and, *mutatis mutandis*, Gärdenfors) showed that indicative conditionals cannot satisfy the Ramsey test and simultaneously have (non-trivial) degrees of acceptance which equal their degree of belief to be true, we assumed with Lewis that the degree of belief of a sentence *A* is derivative to the degree of belief of the proposition $[A]$ that is expressed by *A*, and that the question 'what proposition $[A]$ does get expressed by *A*?' can be answered completely independently of the question 'what is the agent's subjective probability measure *P* like?'. But what if indicative conditionals are sentences with a tacitly *indexical* component? For instance, what if $A \rightarrow B$ really says something like '*My present* subjective probability in *B* given *A* is high'? Then $A \rightarrow B$ would in fact express a proposition *relative to a context by which some probability measure would be supplied*, and relative to two distinct probability measures the conditional might well express *two distinct propositions*; hence, the proposition $[A]_P$ that would be expressed by *A* (relative to *P*) would not be independent of the agent's subjective probability measure *P* anymore, and Lewis' theorem from above would not apply. Indeed, B. van Fraassen (1976) once proved that the Probabilistic Ramsey Test II (Stalnaker's thesis) *is* consistent with the axioms of probability and non-triviality assumptions, as long as one and the same conditional is allowed to express different propositions relative to different probabilistic contexts. But there is also a downside to this result: it follows from an observation by A. Hájek (1989) (see section 31 of Bennett (2003) for a summary) that even this context-sensitive version of the Probabilistic Ramsey Test II can only apply to a non-trivial probability measure, if the probability measure in question is defined on an *infinite* algebra of propositions—no finite non-trivial probability space would do. But how plausible is it to assume that any rational agent who is able to determine one's degrees of acceptability for indicative conditionals by means of a Ramsey test procedure would have to possess infinite powers by which she could discriminate between infinitely many propositions and reason with them? Even if we, *human* agents, do have such powers, it does seem odd to believe that we actually need them in order to determine the degrees of acceptability for even quite simple conditionals. Hence, Hájek's observation just by itself puts so much pressure even on a context-sensitive version of Stalnaker's thesis that the

likeliest option still seems to be that indicative conditionals do not even express propositions relative to a probabilistic context.

So what else could one mean by an assertion of $A \rightarrow B$? The Suppositional Theory suggests to understand indicative conditionals along the lines of other linguistic expressions which do not express propositions but which have a communicative purpose nevertheless. Take 'Yippee!': if I am uttering this, then I am expressing my happy state of mind, without actually saying *that* I am happy. In contrast, the sentence 'I am happy' does express a proposition and therefore has a truth value once the value of 'I' has been supplied contextually. Although the communicative purpose of my utterance of 'Yippee!' is to make you understand that I am happy, I am still able to invite you to update your beliefs about my present emotional state accordingly without asserting any proposition. Moral Expressivism in metaethics holds a similar view about moral statements: they are taken to serve to express some of our attitudes without expressing that we have these attitudes, and for that reason they lack truth values again. Communicating moral statements in this sense aims to affect some other subjects' behaviour, but not by saying that some state of affairs is so-and-so. (Section 42 in Bennett (2003) presents this analogy in more detail.)

Now let us apply a similar line of reasoning to $A \rightarrow B$: assume that any assertion of $A \rightarrow B$ by an agent S expresses (or signals or indicates) S's high subjective probability of B given A without expressing *that* this probability is high. One might say: asserting $A \rightarrow B$ expresses something *de re*—namely, *a high conditional probability*—where an assertion of the sentence '$P(B \mid A)$ is high' would express something *de dicto*—namely, *that the conditional probability in question is high*. And if S asserts $A \rightarrow B$ in the former *de re* sense, then her communicative aim is to alter (if necessary) any *receiver's* subjective conditional probability of B given A: by expressing her own high conditional probability in B given A she intends to change the receiver's conditional probability in B given A so that it becomes high, too. And this can be done, or so the Suppositional Theory has it, without expressing a proposition. (By this we do not mean to suggest that the understanding of 'Yippee!' along expressivist lines necessitates any metacognitive capacities on the sides of the sender or the receiver. We will only argue that this is so for indicative conditionals.)

This suppositional theory of indicative conditionals has a lot of explanatory power:

- It is based on the Ramsey test for conditionals that ties the acceptance of conditionals to suppositional reasoning, which is plausible at least prima facie.
- It steers clear of the Lewisian impossibility result, since according to it, indicative conditionals do not express propositions, which is why degrees of acceptability for indicative conditionals are not degrees of belief-to-be-true.
- It can be made precise in terms of the important scientific concept of subjective conditional probability which makes it part of the philosophically and empirically successful framework of Bayesianism.
- It explains why many indicative conditionals cannot be nested freely and why propositional operators cannot be applied to them unrestrictedly—for once it is accepted that indicative conditionals do not express propositions, then it becomes doubtful whether nesting them or applying propositional operators to them would even be meaningful operations at all. For instance, many instances of conditionals of the form $(A \rightarrow B) \rightarrow C$ are notoriously hard to understanding, which has an obvious explanation if $A \rightarrow B$ does not express a proposition and hence cannot be supposed to be true in a Ramsey-test-like procedure in any obvious manner.
- It supports an elegant and independently justified logical system for indicative conditionals (Adams' conditional logic: cf. Adams 1975).

- It has some empirical support in terms of psychological experiments (see, e.g. Pfeifer and Kleiter 2010): if subjects are given toy stories and then get asked to infer the probabilities of certain conclusions, then—given the assumption that the degrees of acceptability of the indicative conditionals that are involved in these toy stories or in the conclusions are the corresponding conditional probabilities—the majority of these subjects are found to give answers which are rational in the sense of corresponding to the normative ideal.

Obviously, this evidence is far from being conclusive—for instance, Adams' logic of (non-nested) conditionals can be supported also by a *truth-conditional* semantics of conditionals—and suppositionalism has its own problems, too, such as explaining our ability of interpreting particular instances of nested conditionals with great ease, or the existence of some empirical work that seems to speak against Adams' version of the Ramsey test (see Douven and Verbrugge 2010). But suppositionalists about indicative conditionals definitely have made a strong case in favour of their theory. We take it to be a plausible working hypothesis at this point that they are right. In the next section we are going to show why this supports our own thesis from the beginning of this paper.

Metacognition and the defence of the thesis

In light of the last two sections, we claim that *accepting an indicative conditional is a metacognitive process that is not metarepresentational.* Supporting this thesis faces the following difficulty: there is no widely accepted definition of 'metacognitive'. To be sure, metacognition is supposed to be cognition about cognition, but the exact definition and scope of 'about' is not clear enough. So the best that we can do at this point is to highlight coincidences between (1) the suppositional theory of indicative conditionals, and (2) existing views on metacognition that are motivated on independent grounds. We will do this by running through a sequence of claims and quotations from J. Proust's (2007), 'Metacognition and Metarepresentation: is a self-directed theory of mind a precondition for metacognition?'. However, we will not be able to discuss her *arguments* for these claims; see Proust (2007) for the details.

Proust puts forward an understanding of the term 'metacognition' that actually *excludes* metarepresentation, where epistemic feelings, such as the feeling of knowing something (the tip-of-the-tongue phenomenon) constitute her paradigm case examples. While one does not necessarily have to follow Proust in taking the meaning of 'metacognition' thus narrowly construed, it will turn out that her account of metacognition parallels the suppositional theory of acceptance for indicative conditionals to a very significant degree.

First of all, Proust argues that:

- Metacognitive engagements are predictive or retrodictive. Prediction and retrodiction are part of a self-directed evaluative process.

She explains this in terms of a comparison with the outcome of a potential course of action that is to be predicted:

> In order to predict whether you can jump over a ditch, for example, you have to simulate, on the basis of the implicit knowledge of your motor ability and the perceptual cues available, whether you find the jump easy or problematic. In such cases, you simply simulate that you jump, i.e. you imagine yourself as jumping over the ditch in front of you … simulating is just running a dynamic motor representation off-line, and obtaining internal positive feedback on this basis. In conceptual terms: the function of simulation is not to represent yourself as doing something; it is rather to prepare to do something, that is, to do it in a pretend mode … What is true for bodily action prediction also holds for mental action (metacognitive) prediction. (Proust 2007, p. 279.)

So here the task that is analogous to the task of a metacognitive engagement is to predict whether one is able to jump over a ditch. And this, Proust argues, can be done in terms of a dynamic motor representation in which the jump is being simulated by means of the same systems that would be involved in an actual jump. But since this is all done off-line, the jump is only being simulated. Once the simulation has been run, its outcomes can be assessed and translated into a corresponding prediction about would happen in an actual jump. Overall this is an evaluative process that is self-directed in the sense of evaluating one's own bodily processes. And metacognitive processes are supposed to be *like* that, except that it is one's own *mental* processes which are to be evaluated, e.g. one's capacity of remembering a particular mental content.

Clearly, this is in almost perfect correspondence with our analysis of the Ramsey test for indicative conditionals as explained in the first section ('Accepting an indicative conditional'): one could say that the Ramsey test aims to predict the degree of belief that would be assigned to the consequent of an indicative conditional were the antecedent of the conditional to come along as a piece of evidence, which is determined in terms of a simulation again in which the agent's normal inferential or revision capacities are put to the test off-line. (Much more on the topic of offline simulation in general and how it relates to metacognition can be found in Nichols et al. 1996.)

Secondly, Proust argues that

- The metacognitive evaluative process is not explainable in first-order terms.

In her own terms:

> [Some have raised the following objection:] To know what your future disposition is, just look at your prior performance: look at the world, not at the self … To respond to the objection, we can use two types of arguments. One type is conceptual.
>
> (7) Knowing (believing) that a reward of probability p is associated with stimulus S is not equivalent to
>
> (8) Knowing (believing) with probability p that a reward is associated with stimulus S.
>
> In addition to the changing world, a distinctive source of uncertainty may be generated in the knower. (Proust 2007, p. 283.)

What we take Proust to say here is that such metacognitive engagements are not merely instances of cognition about the external world: the probabilities that are the outcomes of metacognitive predictions are not objective, non-epistemic probabilities—objective chances of some event happening—but really subjective probabilities that reflect the agent's uncertainty about what is going to happen. In the case of the Ramsey test, we have made a similar observation: in the Probabilistic Ramsey Test, as explained in the first section ('Accepting an indicative conditional'), the probability to be determined is the subjective conditional probability of the consequent given the antecedent.

However, what does not get perfectly clear just from the earlier quote is why the mental state that is the outcome of a metacognitive act in Proust's sense could not simply be the agents having a particular subjective probability in a proposition about the world, which would still be a first-order mental state (one that would be quantitative in nature). Proust gives further arguments to the effect that this could not be the case. In our own case, we can draw the corresponding conclusion on grounds of a proper theorem: by the impossibility results that we discussed in our second section ('The impossibility result and its consequences'), we can exclude the degree of acceptability of an indicative conditional—the outcome of the Ramsey test—to be a subjective probability in a proposition about the external world, for it would simply be *inconsistent* to assume this to be the case given the Probabilistic Ramsey Test and a non-triviality assumption. The acceptance of an indicative conditional simply *could* not be a first-order process.

Thirdly, Proust continues by pointing out that

- The metacognitive evaluative process is not explainable in second-order terms either.

That is also the sense in which Proust regards metacognition as *not* being metarepresentational. In her own words again:

> Metacognition is not metarepresentational in the sense that there is no 'report' relation between command and monitoring, but a functional complementarity of a basic kind … In each metacognitive intervention, a command token inquires whether typical conditions for a desirable/undesirable outcome (learning, remembering, vs. forgetting, confusing, etc.) are 'now' present; the corresponding monitoring token uses present reafferences to offer a context-based answer (feelings of learning, feelings of knowing, cue-based inferences, etc.) (Proust 2007, p. 285.)
>
> In [the metarepresentational examples] (1) and (11), a thought is represented along with the concept of the propositional attitude in which it is embedded. In contrast, in [the metacognitive example] (12), remembering does not have to be conceptually represented; it only has to be exercised as a trying, that is, simulated … (Proust 2007, p. 286).

And she concludes with:

> In summary: metacognition is neither first-order, nor second-order. We might call this initial, emergent metacognitive level 'level 1.5'. (Proust 2007, p. 287.)

So metacognition just by itself does not involve anything like an agent having beliefs about her own beliefs or similar second-order states or processes: the feeling *of knowing something* differs from believing *that one knows something* in the way that no mental concepts need to be represented in order for the former to take place, but rather that what would be expressed by such mental concepts (such as, *knowing*) is being exercised in a pretend-mode. Accordingly, as pointed out in the first section, the Ramsey test for indicative conditionals exercises an agent's inferential or belief revision capacities in a simulation mode in which it is pretended that the antecedent of the conditional in question is actual evidence; once again no concepts such as *belief* or *evidence* or *inference* need to be represented in the course of that simulation (except, possibly, if they occur explicitly in the antecedent or the consequent of the conditional). And we concluded from the impossibility results in the second section that the degree of acceptability of an indicative conditional could not be a degree of belief in a sentence that would express an introspective proposition, nor a degree of belief in a sentence that would express a proposition relative to a context which would be given by the agent's own subjective probability measure. So accepting an indicative conditional could not be a second-order process either, which is in line with Proust's conclusions.

There are further coincidences. For example, Proust explains why we have difficulties understanding multiple embeddings of metacognitive sentential operators (if she is right, already the case of *I know that I know that I know that A* is problematic; cf. Proust 2007, p. 276), which is in agreement with the difficulties of understanding some instances of nested indicative conditionals (e.g. the case of $(A \rightarrow B) \rightarrow C$ that was mentioned before) and with the suppositionalist explanation of this phenomenon by pointing out that indicative conditionals do not express propositions. If, as Proust argues, introspective knowledge usually does not involve metarepresentations, then, just as understanding iterations of conditional operators is problematic according to the suppositional theory of conditionals, understanding iterations of introspective knowledge ascriptions ought to be problematic according to Proust's theory, and indeed it is. And so forth. But we will leave the discussion at this point, hoping that we have made sufficiently clear why *accepting an indicative conditional* as explained in the first two sections of this chapter *is a metacognitive process that is not metarepresentational* in Proust's sense as explained in this section.

Open questions

We have argued that accepting an indicative conditional seems to be a very good candidate for a metacognitive process in the sense of Proust (2007). But along the way some questions had to be left open which ought to be reconsidered in a more comprehensive treatment of the subject.

Most urgently:

- How can the term 'metacognitive' be explicated so that instances of metacognition that are not metarepresentational are not excluded already on conceptual grounds?

Here the case study of the acceptance of indicative conditionals might actually give us a hint at where to search for an answer. Obviously, in order to answer this question the crucial point will be to explain in what sense metacognition can be cognition *about* cognition without necessarily being an instance of a second-order state of mind, such as an agent's believing *that* she believes that something is the case. And as we said at the end of section 2, suppositionalists about indicative conditionals seem to have an answer: in their view, indicative conditionals express an agent's state of mind without saying that the agent's state of mind is so-and-so. Accordingly, one might expect an explication of 'metacognitive' to proceed in the way that 'cognition about cognition' should at least leave open—or maybe, in Proust's terminology, even *demand*—that metacognitive states and processes are *about* an agent's internal states or processes in the sense of *expressing* these states or processes, in the same sense in which 'Yippie!' expresses a positive emotional state, without stating that the state is so-and-so. Of course, the exact meaning of 'expressing a mental state' needs to be made much clearer itself in order for such an explication of 'metacognition' to succeed, but the merits of the Suppositional Theory of conditionals gives us reason to believe that this might well be doable.

The second open question concerns the Ramsey test for indicative conditionals: in the first section we treated the Ramsey test as if this were the only possible manner in which an agent might come to accept an indicative conditional; and indeed it is treated as such also in the typical expositions of the Suppositional Theory of indicative conditionals. But strictly speaking this cannot be quite right: for instance, if someone whom I take to be competent and trustworthy asserts $A \rightarrow B$, then I might come to accept $A \rightarrow B$ merely on the basis of that assertion, not on the basis of running through the Ramsey test. Presumably, the proper thesis about the Ramsey test and its role in the acceptance of indicative conditionals should not be that the Ramsey test is the *only* manner in which agents may determine the acceptability of indicative conditionals, but that it is the *primary* manner: all other ways of determining the acceptability of such conditionals would derive from it. For instance, in the case where someone asserts $A \rightarrow B$ and I start to accept the conditional as a consequence of that assertion, presumably, the other agent has in fact run the Ramsey test for $A \rightarrow B$, or his acceptance of $A \rightarrow B$ is itself the indirect effect of someone else applying the Ramsey test. But this has to be investigated in more detail. So let us put on record:

- Are all ways of accepting indicative conditionals derivative from some application of the Ramsey test?

When this question gets addressed, it might also turn out that some indirect ways of accepting indicative conditionals do in fact involve second-order mental states; if so, then only the primary way of accepting an indicative conditional—the Ramsey test—could be claimed to be metacognitive in Proust's (2007) sense.

Furthermore, quite obviously, one wonders whether a similar thesis such as the main of thesis of this paper could be defended also for subjunctive conditional, and if yes, what the Ramsey test for subjunctive conditionals would have to look like:

- Is accepting a subjunctive conditional a metacognitive process that is not metarepresentational?

Finally, we are left with a fundamental methodological question. In section 1 and 2 of this paper we presupposed that the agents that we are dealing with were *rational*: they accept indicative conditionals in terms of the Ramsey test because it is rational for them to do so; and they distribute their degrees of belief in agreement with some probability measure for the same reason. But Proust (2007) is actually concerned with real-world agents, that is, concrete humans or animals. So the obvious question arises:

- Is it justified to extend the findings of this paper, which apply to rational agents in the sense explained before, to real-world agents who are not necessarily rational in the same sense?

Clearly, if such a transition is possible at all, it needs some careful argumentation.

There are many more important questions which we did not even formulate, e.g. concerning the implementation of acceptance routines for conditionals in computers, or regarding the relevance of responses to Lewis' triviality results that we did not discuss, such as three-valued semantics for conditionals, or about ways of extending the present thesis on the acceptance of conditionals to further attitudes and mental processes. In any case, we hope this paper constitutes at least some progress in our understanding of the acceptance of indicative conditionals from the viewpoint of the study of metacognition in cognitive science, as well as some progress in our understanding of metacognition from the viewpoint of the philosophy of language and probabilistic accounts of conditionals. But in view of the open questions that we have ended up with in this final section, the present paper is still not more than just a précis of a more elaborate theory of metacognition and the acceptance of conditionals that is yet to be developed.

Acknowledgements

This work was supported by the AHRC, the ESF, and the Alexander von Humboldt Foundation.

References

Adams, E. W. (1975). *The Logic of Conditionals*. Dordrecht: Reidel.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.

Douven, I. and Verbrugge, S. (2010). The Adams family. *Cognition*, 117, 302–18.

Gärdenfors, P. (1986). Belief revisions and the Ramsey test for conditionals. *Philosophical Review*, 95, 81–93.

Gärdenfors, P. (1988). *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Cambridge, MA: MIT Press/Bradford Books.

Hájek, A. (1989). Probabilities of conditionals – revisited. *Journal of Philosophical Logic*, 18(4), 423–8.

Leitgeb, H. (2007). Beliefs in conditionals vs. conditional beliefs. *Topoi*, 26(1), 115–32.

Leitgeb, H. (2010). On the Ramsey test without triviality. *Notre Dame Journal of Formal Logic*, 51(1), 21–54.

Leitgeb, H. (2011). Ramsey = God – Moore. (A reply to Chalmers and Hájek). *Topoi*, 30(1), 47–51.

Leitgeb, H. (2012). A probabilistic semantics for counterfactuals. Part A. *Review of Symbolic Logic*, 5(1), 26–84.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, 85, 297–315.

Nichols, S., Stich, S., Leslie, A., and Klein, D. (1996). Varieties of off-line simulation. In P. Carruthers and P. Smith (Eds.) *Theories of Theories of Mind*, pp. 39–74.Cambridge: Cambridge University Press.

Pfeifer, N. and Kleiter, G. D. (2010). The conditional in mental probability logic. In M. Oaksford and N. Chater (Eds.) *Cognition and Conditionals: Probability and Logic in Human Thought*, pp.153–73. Oxford: Oxford University Press.

Proust, J. (2007). Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition?. *Synthese*, 159, 271–95.

Ramsey, F. P. (1929). General propositions and causality. In D. H. Mellor (Ed.) *F. P. Ramsey: Philosophical Papers*, pp.145–63. Cambridge: Cambridge University Press (1990).

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.) *Studies in Logical Theory*, pp. 98–112. Oxford: Basil Blackwell.

Stalnaker, R. (1970). Probability and conditionals. *Philosophy of Science*, 37, 64–80.

Van Fraassen, B. (1976). Probabilities of conditionals. In W. Harper and C. Hooker (Eds.) *Foundations of Probability theory, Statistical Inference, and Statistical Theories of Science*, Volume I, pp. 261–308. Dordrecht: Reidel.

Chapter 16

Is hypnotic responding the strategic relinquishment of metacognition?

Zoltán Dienes

In this chapter the importance of understanding metacognition at the level of self-ascription of mental states (to use the terminology of Proust (forthcoming)) will be highlighted. I argue that hypnosis might provide a sort of show case for metacognition researchers, because a mere change in ascriptive metacognition results in behaviour so bizarre that many people believe that response to hypnotic suggestion involves either faking or else an extraordinary change in first-order states (attention or other abilities). Hypnosis illustrates the dramatic effect a small change in metacognition can have, even on such an everyday activity as raising one's arm. Consider the hypnotic suggestion that the arm will rise by itself, and the person looks in amazement at their rising arm. Is the person faking their claims of involuntariness? Or can they selectively attend in such a remarkable way they inhibit all information that contradicts the hypnotic suggestion, allowing the suggested response to happen? I argue that neither faking nor first-order attentional abilities are typically involved, but rather metacognition.

The cold control theory of Dienes and Perner (2007; see also Barnier et al. 2008) takes a common component from previous theories in hypnosis (e.g. the sociocognitive tradition represented by Spanos (1986), and the normally contrasting neo-dissociation theory of Hilgard (1977)) and identifies it as the essence of hypnosis: Namely, hypnotic response is constituted by intending to perform some motor or cognitive action, while remaining unaware of the intention—in fact, the hypnotized subject actively thinks she is not intending to perform the action. Construed in this way, hypnosis is a purely metacognitive phenomenon. It involves no changes in first-order abilities, i.e. abilities that rely on mental states that are only about the world. If one intends to lift one's arm it will rise; but if one is resolutely unaware of the intention, the arm will appear to lift by itself, producing the phenomenology of hypnosis. Cold control theory claims that there is nothing more to hypnotic responding than the metacognitive change. The ability to raise one's arm is unexceptionable; what is strange in hypnotic response is only the metacognition that goes with it. (Cold control is 'cold' because it is executive control without an accurate higher-order thought (HOT): control without the HOT). On this theory, incidentally, it follows that no animal could be hypnotized unless it had mental state concepts, at least of intentions, so that it could intentionally do something while believing it wasn't intentional!

This chapter will explore evidence and predictions of this metacognitive approach to understanding hypnosis. Initially, I will indicate what I take the phenomena of hypnosis to be that I want to explain. Then, the term metacognition will be defined as it is used in this chapter. Next, the chapter will consider the relation between hypnosis and metacognition. First it overviews the correlates of hypnotizability and suggests metacognitive correlates may be more promising than those measuring first-order cognitive abilities. Then, data are described investigating the effect on hypnotic response of impairing prefrontal regions previously shown to be involved in establishing

accurate higher-order thoughts. Finally the prediction of the theory that no first-order abilities are acquired by virtue of responding hypnotically (the only change is metacognitive) is considered, and relevant empirical evidence discussed for the absence of hypnotically-induced first-order abilities.

What is hypnosis?

Hypnosis is a situation in which a person creates altered experiences of volition or reality in accord with the requirements of the situation (see Nash and Barnier (2008) for recent reviews of the field). The requirements can be provided by the suggestions of another person, for example a hypnotist, but they can also be produced by the goals of the person themselves. An example of an altered experience of volition is 'magnetic hands', which 90% of people can experience: hold your hands out in front of you palms facing each other and imagine your hands are magnets, creating a force pulling your hands together. You have responded successfully if you feel a force pulling your hands together, as if by themselves. If you just tried this now, the goal was set by yourself without the need for a hypnotist to be present. An example of an altered experience of reality is hallucination. Look at a pale object in front of you and make it look red. If you can experience the redness as genuinely external and in the object, that is if it seems like you are seeing it as red, you have responded successfully. Only about 10% of people can reliably respond to such cognitive suggestions. Alterations in the experience of reality can also have delusional components, for example, it can be suggested that people have changed gender (McConkey et al. 2001). There is a normal distribution of response to hypnotic suggestion with the top 10% of responders called 'highs' and the bottom 10% 'lows'; those in between are 'mediums'. Why some people are more hypnotizable than others remains unresolved (Heap et al. 2004).

The English word 'hypnosis' suggests a state—in everyday speak we can say a person can enter hypnosis, hypnosis can be induced, and a person may be hypnotized. In the definition given I have deliberately avoided having a special state be constitutional of hypnotic responding: I take the role of a special state to be a theoretical and empirical issue that could go in several directions, with the fundamental phenomenon we trying to explain (the ability of some people to alter the experience of volition and reality) remaining the same (cf. Kirsch et al. 2011). But some facts are worth bearing in mind. There is a relatively standard procedure that serves as a 'hypnotic induction' consisting of suggestions for relaxation and sleepiness. Often a condition is operationally defined as hypnotic in experimental research if it was preceded by this induction. Yet the induction causes only a small increase in the rate of responding to suggestions. For example, if seven standard suggestions are given without a hypnotic induction and on average two are passed, an induction will increase the response rate to 2.5 (Braffman and Kirsch 1999). Further, this increase in response can be accounted for by an increase in expectancy of responding. Correspondingly, almost any procedure can serve as an induction, including: stationary cycling with suggestions of becoming more alert; a sugar pill labelled 'hypnotic'; inert medically scented gas; stroking the head; staring in the eyes; pressure on the thumbs; drinking magnetized water; a blow on a gong; or the simple 'now you are hypnotized' (Lynn et al. 2008). There may or may not be a special state that slightly facilitates hypnotic response (see Oakley (2008) for a recent review), but the aim of this chapter is not to focus on explaining any such state. Cold control might be facilitated in a special state, or the special state could just be another suggestion implemented by cold control.

One theory of hypnotic response that has to be considered first, even if just to dismiss, is that hypnotic response is faked. However, at least in academic research settings, highs, unlike people asked to fake being high, carry on responding to suggestions even when they think they are alone unobserved (Kirsch et al. 1989); highs, unlike people asked to fake being highs, pass lie detector tests of honesty (Kinnunen et al. 1994); and highs responding to suggestion rather than faking

response have different brain regions involved, with the brain regions being consistent with those expected if the responses were veridical (Ward et al. 2003; Oakley 2008).

Two broad approaches have historically been used to explain hypnosis. One explains the responses as purely responses to demand characteristics according to well-established, everyday psychological principles (e.g. Spanos 1986). Social pressure has a powerful effect on people's actions. If the two people either side of you on stage are acting like chickens, you can either act like a chicken or ruin the show. Before you know it, you are acting like a chicken. Such an explanation can be combined with a metacognitive approach like cold control to argue that demand characteristics lead people to be unaware of their intentions. But it need not. Demand characteristics might create expectations, and the expectations might directly cause the hypnotic experiences, just as expectations directly cause placebo effects (Kirsch 1985). In this case no further metacognitive explanation is necessary. If a person believes they did not intend the mental or physical action, they are just being accurate.

The dissociation approach explains hypnotic response as a result of a change in cognitive control structures, with one structure splitting (dissociating) either within itself or from the other structures (Hilgard 1977). Such a change could involve changes in awareness of intentions, for example, if the central executive split in two and what one half intended the other half was not aware of, as required by cold control. But dissociation theory does not require this. Dissociation might split off a control structure from the central executive, so its action were no longer triggered by intentions but by hypnotic suggestions (Bowers and Woody 1996). In this case also no further metacognitive explanation is necessary. If a person believes they did not intend the mental or physical action, they are just being accurate.

In sum, cold control theory is not an alternative theory to the main theories, it is a way of thinking about each. But it can be proved wrong, as we will see.

What is metacognition?

Metacognition is most broadly construed cognition about cognition. Cognition can be conceptual or non-conceptual; thus cognition about cognition can occur conceptually or non-conceptually. For example, Shea (in press) argues that the error signal in connectionist networks has representational content that is about the accuracy of the connectionist representation: Buried deep at the level of small numbers of neurons, any physiological error-signal is non-conceptually about non-conceptual representation, yet in the end about the accuracy of mental states a person may be in. At the other end, Rosenthal (2005) discusses conceptual thoughts about whether one is in a certain mental state. Proust (forthcoming) calls the latter metacognition 'ascriptive', a general process of conceptually representing one's mental states, a process which could be applied to other people's mental states as much as one's own. She urges the term 'metacognition' be reserved for the non-conceptual abilities dedicated to evaluating one's own mental dispositions, as shown for example by the ability of some animals to evaluate their own cognitive accuracy without being able to pass theory of mind tasks about other individuals. The reader should bear in mind I am using metacognition to talk about the higher-order thoughts of Rosenthal; the term can be substituted in the reader's mind with another if they have another preferred term for ascriptive metacognition.

Now I will consider the evidence for a link between metacognition and hypnosis.

Correlates of hypnotizability

According to one line of thinking, highly hypnotizable people are skilled in sustaining attention (Crawford et al. 1993) perhaps especially in inhibiting distracting or contradictory information in

the world (e.g. David and Brown 2002). This is a first-order skill in so far as it involves ability to attend to the world or ignore distractions in the world. The relation between inhibitory ability without a hypnotic induction and hypnotic suggestibility has been studied most directly using the Stroop effect and negative priming. Studies using the Stroop test have produced conflicting findings, with either no difference between highs and lows or with differences in either direction.[1] A further way of assessing cognitive inhibition is with a negative priming task, in which participants are instructed to attend to some stimuli and ignore others. Dienes et al. (2009) found with 180 participants the correlations between hypnotizability and negative priming or between hypnotizability and latent inhibition were close to zero, with upper limits of about 0.20. Similarly, Varga et al. (2011) with 116 subjects found no significant correlations between hypnotizability and reaction time measures of sustained, selective, divided or executive attention.

In sum, there is no clear relation between hypnotizability and ability to inhibit information. If hypnotizability is related purely to those individual differences that exist between adult humans in metacognitive processes, these null results are to be expected. However, there is a striking exception to this overall conclusion, based on the work of Raz and his group. When highs are given the suggestion that words will appear to them as meaningless, the Stroop effect can be substantially reduced (e.g. Raz et al. 2002, 2003; see also Ianni et al. 2006). The suggestion is just as effective whether or not a hypnotic induction is given (Raz et al. 2006), so appears not to depend on being in a special state, but on having a certain ability. The effect appears non-existent to weak in lows (Raz and Campbell 2011). In as yet unpublished studies Ben Parris and I have also found the effect of this suggestion significantly less in lows than highs, even when the context is not defined to subjects as hypnotic i.e. the suggestion is given as an exercise in imagination and no induction is used: In this context, lows should not be motivated to perform badly. Note however the response was still hypnotic for highs in the sense that they produced altered experiences of reality. In sum, there appears to be an individual difference ability in reducing the effect of conflicting information, where highs can overcome conflict by use of imagination but lows cannot. It is intriguing how this can be reconciled with the equivalent uninstructed performance of highs and lows on the Stroop. Could highs be able to generate especially vivid images, overwriting the contents of perception? Yet on standard paper and pencil ratings of vividness of imagery, there is little to no relationship with hypnotizability (see Jamieson and Sheehan 2002). The relation between hypnosis and a reported tendency to imaginative absorption has long been noted (e.g. Roche and McConkey 1990; Nadon et al. 1991) though what abilities this entails is less clear (Jamieson and Sheehan, 2002). I will discuss the Raz effect further later; for the time being, just note that the ability to overcome Stroop in a certain context is a phenomenon for future research to attack the metacognitive approach to hypnosis at its weakest, because it appears to be a case where hypnotic response involves having a special ability not available non-hypnotically.

[1] Without hypnotic induction or suggestions being used, most studies have found no significant difference between highs and lows on Stroop interference (Kaiser et al. 1997; Aikens and Ray 2001; Kallio et al. 2001; Egner et al. 2005). Dixon et al. (1990) and Dixon and Laurence (1992) found significantly more Stroop interference in highs than lows; however, Rubichi et al. (2005) found significantly less Stroop interference in highs rather than lows. On a related task, Iani et al. (2006) found that highs and lows without an induction were not detectably different in terms of the effect of irrelevant flanking items on the classification of a central letter. Farvolden and Woody (2004) tested proactive interference in highs and lows. Participants were trained on one set of paired associates (AB) then on three study-test trials of a second set (AC). On the first test trial of the second set, highs made more errors in recalling C to the cue A than lows did. Thus, highs may have found it harder to inhibit the effect of the first set of words, which is not consistent with highs being good at inhibition.

It will be useful to dismiss a theory some people may have about hypnosis that motivates the plausibility of a link between attention/inhibitory ability and hypnotizability. The theory that goes back to James Braid in 1847 and was revived by Baars (1988) is that successful hypnotic response occurs because highs maintain a persistent uncontradicted image of the required result. To test the theory, Zamansky and Clark (1986) asked subjects to engage in imagery inconsistent with the hypnotic suggestions given (e.g. for a rigid arm suggestion, to imagine a different world in which their arm is bending). Highs were just as responsive to suggestions (e.g. that the arm is unbendable) when engaged in imagery inconsistent with the suggestion as when having consistent imagery, even as they concurrently reported the imagery. That is, their arm remained unbent, even as the subjects described an image of the arm bending. Thus, the theory that highs attend to one idea and inhibit all else in order to achieve hypnotic response is false.

Given that first-order abilities are similar for high- and low-hypnotizable subjects, cold control theory indicates that what is important is to assess individual differences in tendency to produce accurate higher-order thoughts. Designing a task to measure second-order (metacognitive) processes without first-order confounds is difficult, as the chapters in this volume on exploring metacognition in animals illustrates (see, e.g. Perner Chapter 6 and Couchman Chapter 1). Fortunately, the evidence for no difference in first-order abilities between highs and lows allows second-order differences to be explored more easily. We have begun exploring a test of second-order thoughts. For example, in unpublished work, Karin Berg at Sussex asked subjects to keep looking at a candle while trying to either (a) remain at all times aware of seeing the candle for 10 minutes (meditation task) or (b) not consciously see the candle for 10 minutes (ignore task; compare Wegner's (1994) 'white bear' ironic control task, where people are asked to not think of a white bear). People were asked at random intervals (roughly once a minute) whether they were just that instant before aware of seeing the candle. Because people remained physically looking at the candle there was a persistent first-order visual representation of the candle; but to what extent did people have accurate higher-order thoughts about seeing the candle? The difference between (a) and (b) in reports of seeing the candle was taken as measuring control in having accurate higher-order thoughts, and the total number of reports of seeing the candle in both (a) and (b) as measuring coupling of higher-order thoughts to first-order states, i.e. the tendency to have an appropriate higher-order thought given that a first-order state exists. Higher-order thought control did not correlate with hypnotizability ($r = -0.23$, ns), but higher-order thought coupling did ($r = -0.54$). That is, highly hypnotizable people appeared generally prone to inaccurate higher-order thoughts (regardless of their intentions): it is not that they have good metacognitive control over higher-order thoughts but that higher-order thought coupling is weak. The relation between HOT coupling and hypnotizability held for each task separately: for the meditation task there was a negative correlation between number of times they were aware of the candle and hypnotizability ($r = -0.47$; contrast Van Nuys 1973); and so was there for the ignore task ($r = -0.46$; cf. Bowers and Woody 1996). (Note in the meditation task, the results show highs failing to follow instructions well and in the ignore task the results show highs following instructions well.) The relation between coupling and hypnotizability held even after partialing out expectation of responding to each suggestion, motivation to respond to each suggestion, and a paper and pencil measure of sensitivity to social desirability (i.e. tendency to say things to please people: Marlow Crowne test, an attribute which was not controlled in the Van Nuys study). The apparent weak coupling may allow highs to decide in appropriate contexts to *forgo* higher-order thoughts of intending in order to respond hypnotically to suggestions. These results are preliminary ($N = 20$), and Rebecca Semmens-Wheeler is following them up at Sussex by using a succession of images rather than a candle to focus on; we can then test the extent to which people were taking in the images in later memory tests to verify reports of consciously thinking of the images or not. While the results are preliminary, at least

cold control suggests a line of enquiry not pursued before in understanding the correlates of hypnotizability: Individual differences in ascriptive metacognition.

Manipulating the neural substrate of metacognition

Now we consider whether we can do something to change hypnotizability, specifically by affecting people's ability to engage in accurate metacognition. Lau and Passingham (2006) found two masking conditions where people could discriminate one of two shapes to an equal degree but the conditions differed in the extent to which people were aware of seeing the shapes rather than thinking they were just guessing. That is, first-order abilities were equivalent, but metacognitive abilities differed. Functional magnetic resonance imaging (fMRI) indicated that a single cortical area distinguished the conditions, the mid dorsolateral prefrontal cortex (DLPFC). Further, when Rounis et al. (2010) disrupted the area with theta burst transcranial magnetic stimulation (TMS), subject's awareness of seeing, as revealed by their ascriptive reports of seeing, was disrupted even when first-order perception was titrated to be the same with and without TMS. That is, the disruption Rounis et al. found was purely metacognitive. If the area is responsible for accurate higher-order thoughts in general, disrupting the region with repetitive TMS (rTMS) or alcohol should make it harder to be aware of, for example, intending to perform an action. That is, it should be easier to subjectively respond to a hypnotic suggestion. Sam Hutton and I, in as yet unpublished work, tested this prediction of cold control theory with TMS. Twenty-four mediums were subject to rTMS to the DLPFC and to a control site, the vertex, in counterbalanced order. The hypnotist was blind to which site had been stimulated. Subjects gave ratings on a 0–5 scale of the extent to which they experienced the response, for four suggestions (magnetic hands, arm levitation, rigid arm, and taste hallucination). Overall, rTMS to the DLPFC rather than vertex increased degree of subjective response by about a third of a rating point on average. Further, subjects did not differ in their expectancy that they would respond in the two conditions, so the rTMS had an effect on hypnotic experience above and beyond expectancies. A further study conceptually replicated the effect, but this time using alcohol. The dorsolateral prefrontal cortex is disrupted by alcohol and surprisingly previous research has not investigated the effect of getting drunk on hypnotizability. Rebecca Semmens-Wheeler, Dora Duka, and I recently explored the effect of alcohol on hypnotic response. Thirty-two mediums were assigned to either an alcohol or placebo alcohol condition; those in the alcohol condition drunk the equivalent of roughly five glasses of wine over a 30-minute period. Both groups were then tested on nine suggestions and various frontal tasks. Alcohol indeed disrupted frontal function. Crucially, alcohol increased subjective response by one scale unit compared to placebo, on the same scale as used in the TMS study. While the prediction of cold control that disruption of the DLPFC would enhance hypnotic response was confirmed in both experiments, both TMS stimulation and alcohol would have affected a large area of prefrontal cortex subserving numerous functions, not just metacognition. Thus the results are also consistent with other theories, such as that of Woody and Sadler (2008) who postulate hypnosis is a state of diminished frontal function. However, the situation is not one of stalemate. Cold control in principle specifies which areas are the important areas for future work, as technology allows more specific areas of the cortex to be targeted.

Hypnotic versus non-hypnotic abilities

According to cold control theory, a person has no first-order abilities in responding to a hypnotic suggestion that they did not have already. The difference is only that performing the action hypnotically makes it feel like it is happening by itself. That is, the only difference between responding hypnotically and non-hypnotically is a metacognitive one in the sense of forming higher-order

thoughts about first-order states. This is a controversial claim. For example, Hilgard (1977) and others have claimed that there is a hypnotic mechanism of pain reduction not available non-hypnotically. On the other hand, those in the sociocognitive tradition (e.g. Spanos 1986) claim people can reduce pain equivalently by hypnotic suggestion as by the use of cognitive strategies: Both techniques essentially involve eliminating 'catastrophizing' cognitions (i.e. thoughts that one is being badly harmed), generating positive thoughts, reinterpreting sensations, and controlling attention. Cold control aligns itself with the sociocognitive tradition in this respect. The difference between cognitive behavioural and hypnotic methods is only that in the former the person is aware of actively engaging in strategies while in the latter the pain seems to go away by itself. According to Spanos, the problem with studies that have found a difference between hypnotic and non-hypnotic suggestions is that when subjects are aware that a hypnotic condition will be compared to a non-hypnotic one, they like to please the experimenter by 'holding back' in the non-hypnotic condition in order that they can perform better in the hypnotic condition. Further, studies comparing hypnotic and non-hypnotic conditions have to control expectancy, as different expectancies could produce different degrees of placebo pain relief (cf. Kirsch et al. 1995). Studies directly comparing hypnotic with cognitive behavioural treatments for experimental pain often have not found differences between these conditions (Milling et al. 2002) even when the authors argue for hypnotic techniques involving a different mechanism (e.g. Miller and Bowers 1986, 1992). On the other hand, Derbyshire et al. (2009), for example, found the same suggestion given with an induction rather than without produced slightly greater degrees of pain relief as revealed in subjective ratings and in the 'pain matrix' (the brain areas involved in pain as revealed in fMRI)—but here hold-back and expectation effects seem likely. So no conclusive answer can be given about whether hypnotic pain relief is more effective than non-hypnotic pain relief. But if there is a difference, it is small. Further, there is less awareness of using cognitive strategies in hypnotic rather non-hypnotic pain relief (Miller and Bowers 1986; Hargadonet al. 1995), just as a metacognitive account predicts.

Pain relief might be regarded as a sort of negative hallucination; it is at least the removal of a perception one would otherwise have. Hallucinations generally might strike the reader as a case where people do something hypnotically they could not do otherwise: Hypnotically people can take themselves as perceiving something they would not perceive otherwise. For example, Kosslyn et al. (2000) argued that people could 'see' colours with hypnotic hallucination that they could not see with imagination. Kosslyn et al. asked highly hypnotizable subjects to either see a colour pattern in grey-scale, or to see a grey-scale pattern in colour. Positron emission tomography scanning indicated that the left and right fusiform areas were active in highs either seeing genuine colour or hallucinating colour, but not when veridically seeing greyscale. When asked to imagine the same colour changes, activation changes were not detected in the left fusiform. In interpreting the latter result, however, one should bear in mind Kosslyn et al's concern that the subjects did not 'drift into hypnosis' and hallucinate in the imagination condition. The wording in the imagination condition was chosen to 'lead the subjects to attend to the visible stimulus and alter it rather than to substitute a complete hallucination'. That is, the demand characteristics entailed forming a less convincing image in the imagination rather than the hallucinate condition. It is thus not surprising that this was reflected in less relevant activity in the fusiform area for the imagination condition than the hallucination condition. Both hold-back and expectation effects are likely to operate. Cold control theory predicts that people will be able to intentionally produce the same vivid experience in imagination as when hallucinating, and to produce the same fusiform activities, with a hypnotic induction being irrelevant. Kirsch et al. (2008) gave subjects exactly the same suggestion with or without induction and subjects rated how much colour they saw on a 0–100% scale. Subjects could drain or add substantial amounts of colour when given a suggestion, and there was no evidence that hypnotic induction made a difference. Further, without

induction, subjects did not rate themselves as hypnotized, so a 'drifting into hypnosis' hypothesis was ruled out. Recall similar results were obtained for the suggestion that words become meaningless: the effect of the suggestion was just as strong whether or not a hypnotic induction was given (Raz et al. 2006).

Further, McGeown et al (in press) repeated the procedure of Kirsch et al with fMRI. Brain imaging confirmed the subjective results: Hypnotic induction improved both subjective response and activation in visual areas by only a small amount (with no difference detected in the left or right fusiform).

We have not quite yet established what we need for these hallucination cases to show that people have the same abilities when responding hypnotically as non-hypnotically. One issue is whether hypnosis involves a special state that can be induced; the other is whether a person responds hypnotically. A person can respond to a suggestion, e.g. creating the feeling of a magnetic force pulling their hands together, or even hallucinating an object change colour, without any special state having been induced (e.g. Braffman and Kirsch 1999). In this sense, hypnosis is a way of doing, not a way of being. Thus, showing that a hypnotic induction is not needed for subjects to experience hallucinations does not yet show that a person can do non-hypnotically whatever they can do hypnotically (cf. Kirsch et al. 2011). So the question is, does responding hypnotically—responding successfully to suggestions for altered perceptions and volition—consist in performing a (bodily or mental) action (no better or worse than one could do normally) (simply) while believing one is not intending it? Responses to cognitive suggestions involve performing mental rather than bodily actions (cf. Proust forthcoming). What is the mental action involved in hallucinating colour? According to cold control it is imagining the colour. Following Frith (1992), if one imagined the colour but was unaware one intended to so imagine, the resulting visual representations are not taken to be self-generated, so therefore they are generated by the world: the subject experiences seeing. The prediction of cold control is that imagination will be just as vivid as hallucination—it is just the former will be taken to be internally generated while the latter appears external. Indeed, hypnotic hallucinations can be flimsy and transparent (McConkey et al. 1991), though a detailed comparison with non-hypnotic imagination is still required (with hold-back and expectation controlled). More problematic is the Raz example, whereby subjects seem able to inhibit the reading of words under suggestion, even though they are not especially good at inhibiting words with no suggestion, under normal Stroop conditions (as discussed earlier). How can cold control explain this fact? It may be that there is a strategy that subjects could implement voluntarily to overcome Stroop: the suggestion implicitly provides it, but subjects do not realize they can use it quite generally (cf. Sheehan et al. 1988). Ben Parris, Lynne Somerville, and I have just started testing subjects by first giving subjects the experience of the suggestion that they cannot read the word, then telling them that they can use this strategy voluntarily at any time in order to overcome Stroop. We are taking ratings of volition, depth of hypnosis, and alterations in perception to determine if highs in no special state can voluntarily reduce the Stroop effect by intentional use of imagination, experienced as imagination. Cold control theory predicts that they can. But if people need to change their experience of specifically *perception* in order to overcome the Stroop effect, then the metacognitive account of hypnosis fails, at least for hallucinations.

Keeping it real

Hypnosis provides an interesting test case for pursuing the distinction Proust (forthcoming) makes between metacognition involved in bodily versus mental actions. For example, bodily actions involve relatively known mechanisms of efferent copy and feedback signals in ways that contribute to the experience of volition. Monitoring one's actions is widely distributed amongst

species, while monitoring cognition is restricted to a few. Thus, action and cognition monitoring may not share mechanisms at a detailed level. Yet Proust argues that mental and bodily actions are at a functional level equivalent; one can intend certain epistemic outcomes, and get feelings (e.g. of knowing) providing feedback on whether the outcome is being successfully approached. It is just such an analysis that cold control requires to be a general theory of hypnosis. If intending to act behaviourally is qualitatively different from intending to act mentally, there is challenge to cold control theory in uniting them in a single account of behavioural and cognitive suggestions. Metzinger (2009) similarly also distinguishes different types of agency—attentional, cognitive, and bodily. Indeed, hypnotic suggestions are often broadly divided by hypnosis researchers into motor (e.g. the suggestion that an arm will rise) and cognitive (e.g. hallucinations, delusions, amnesia). Negative hallucinations likely involve altering attention (e.g. away from painful stimuli). Is there a single mechanism at a broad enough level of description to unite the different actions and agencies, as Proust suggests? Motor suggestions are on average easier than cognitive ones, with 90% of people able to pass one or other motor suggestion without being able to pass a cognitive one. So the divide is real at one level. On the one hand, the rTMS study reported earlier involved changes to both motor and cognitive suggestions. On the other hand, whether cold control can explain the sense of 'reality' of hallucinations and the conviction of delusions remains a possible weak point in the theory, with the theory fracturing precisely down the groove between different meta-entities that Proust and Metzinger identify. In the Kirsch et al. (2008) study subjects were asked to 'make the display coloured' or to 'drain the colour': Responding to the suggestion involved intentionally doing something where the intention could be conscious. Further, in Derbyshire et al. (2009) subjects rated more control over pain when given a suggestion after a hypnotic induction rather than without. So can hallucination really be based on being unaware of intentions? As Proust discusses, a mental action is an action because it partakes in a chain of mental events with a specified epistemic goal. The epistemic goal could be to have a certain perception. This can be achieved by a number of actions only some of which are intended but without awareness of so doing. For example, in the Derbyshire et al. study, subjects imagined turning a dial in order to change the level of pain; conscious intentional control was exerted in changing the imagery of the dial—the link from the dial to pain relief need not occur by strategy of which the subject is conscious of intending. Thus, a subject may experience themselves as intentionally changing perception, even while cold control remains the mechanism by which imagination becomes mistaken as perception. The detailed phenomenology required by cold control will be interesting to test in future studies. Further, future research should investigate the non-conceptual underpinnings of the ascriptive metacognitive changes cold control postulates.

Proust's (forthcoming) analysis of mental and bodily acts being functionally equivalent at a broad enough level of description also facilitates another metacognitive theory of hypnosis, different from cold control: The discrepancy-attribution theory of Barnier and Mitchell (see Barnier et al. 2008). On this view, a hypnotic response involves evaluating the effort produced in a mental or bodily act—if the mental act is surprisingly easy, an explanation is sought in terms of an external cause, leading to attributions of involuntariness (in the case of motor suggestions) or perception (in the case of hallucinations). See Barnier et al. (2008) for a comparison of the two theories.

The link between intention and hallucination used in cold control was first postulated by Frith (1992) and Bentall (1990) in order to explain schizophrenia (though later dropped by Frith). Yet in some ways schizophrenia and hypnotic response are opposites. In schizophrenia, hallucinations happen in ways detrimental to the person's overall goals. In hypnosis, response occurs to further the person's goals and hence is appropriate for the context. For example, a post-hypnotic suggestion to touch one's eyebrow when the word experiment is mentioned is not elicited in a different context from which it was given (Spanos et al. 1987). Further, hypnosis appears to have

no special compulsive power to make people perform antisocial acts against their principles (Coe et al. 1973). Indeed, people with schizophrenia score below average on hypnotizability (see Pettinati (1982), for a critical review). That is, the relinquishment of metacognition in hypnosis is strategic and specific, unlike in schizophrenia.

Because cold control is used in the service of overall goals, it can be placed in an evolutionary context. Whatever selective forces resulted in people acquiring ascriptive metacognitive abilities (and they are surprisingly hard to specify: Rosenthal 2008), there may be a selective reason for people to strategically remain unaware of their intentions in certain contexts. Dienes and Perner (2007) argue that cold control has shown itself in every continent through all known history—in the form of spirit possession. Not only is spirit possession widespread it comes with certain advantages when it is genuinely contextually appropriate and involves genuine self deception (i.e. when it involves cold control).

In sum, I argue hypnosis is quintessentially an alteration in metacognition, and both hypnosis and metacognition researchers would benefit from working together to understand its nature.

References

Aikens, D. and Ray, W. J. (2001). Frontal lobe contributions to hypnotic susceptibility: A neuropsychological screening of executive function. *International Journal of Clinical and Experimental Hypnosis*, 49, 320–9.

Baars, B. (1988). *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

Barnier, A. J., Dienes, Z., and Mitchell, C. J. (2008). How hypnosis happens: New cognitive theories of hypnotic responding. In M. Nash and A. Barnier (Eds.) *The Oxford Handbook of Hypnosis: Theory, Research, and Practice*, pp. 141–178. Oxford: Oxford University Press.

Bentall, R. P. (1990). The illusion of reality: a review and integration of psychological research on hallucinations. *Psychological Bulletin*, 107, 82–95.

Bowers, K. S. and Woody, E. Z. (1996). Hypnotic amnesia and the paradox of intentional forgetting. *Journal of Abnormal Psychology*, 105, 381–9.

Braffman, W. and Kirsch, I. (1999). Imaginative suggestibility and hypnotizability: An empirical analysis. *Journal of Personality and Social Psychology*, 77, 578–87.

Coe, W. C., Kobayashi, K., and Howard, M. L. (1973). Experimental and ethical problems in evaluating the influence of hypnosis in antisocial conduct. *Journal of Abnormal Psychology*, 82, 476–82.

Crawford, H. J., Brown, A. M., and Moon, C. E. (1993). Sustained attentional and disattentional abilities: Difference between low and highly hypnotisable persons. *Journal of Abnormal Psychology*, 102, 534–43.

David, D. and Brown, R. J. (2002). Suggestibility and negative priming: Two replication studies. *International Journal of Clinical and Experimental Hypnosis*, 50, 215–28.

Derbyshire, S. W. G., Whalley, M. G., and Oakley, D. A. (2009). Fibromyalgia pain and its modulation by hypnotic and non-hypnotic suggestion: An fMRI analysis. *European Journal of Pain*, 13, 542–50.

Dienes, Z. and Perner, J. (2007). The cold control theory of hypnosis. In G. Jamieson (Ed.) *Hypnosis and conscious states: The cognitive neuroscience perspective*, pp. 293–314. Oxford: Oxford University Press.

Dienes, Z., Brown, E., Hutton, S., Kirsch, I., Mazzoni, G., and Wright, D. B. (2009). Hypnotic suggestibility, cognitive inhibition, and dissociation. *Consciousness & Cognition*, 18, 837–47.

Dixon, M. and Laurence, J. -R. (1992). Hypnotic susceptibility and verbal automaticity: Automatic and strategic processing differences in the Stroop-colour naming task. *Journal of Abnormal Psychology*, 101, 344–7.

Dixon, M., Brunet, A., and Laurence, J. -R. (1990). Hypnotizability and automaticity: Toward a parallel distributed processing model of hypnotic responding. *Journal of Abnormal Psychology*, 99, 336–43.

Egner, T., Jamieson, G., and Gruzelier, J. (2005). Hypnosis decouples cognitive control from conflict monitoring processes of the frontal lobe. *Neuroimage*, 27, 969–78.

Farvolden, P. and Woody, E. Z. (2004). Hypnosis, memory, and frontal executive functioning. *International Journal of Clinical and Experimental Hypnosis*, 52, 3–26.

Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia*. Hove: Psychology Press.

Hargadon, R., Bowers, K. S., and Woody, E. Z. (1995). Does counterpain imagery mediate hypnotic analgesia? *Journal of Abnormal Psychology*, 104, 508–16.

Heap, M., Brown, R. J., and Oakley, D. A. (Eds.) (2004). *The highly hypnotisable person: Theoretical, experimental, and clinical issues*. London: Routledge.

Hilgard, E. R. (1977). *Divided consciousness: Multiple controls in human thought and action*. New York: Wiley-Interscience.

Iani, C., Ricci, F., Gherri, E., and Rubichi, S. (2006). Hypnotic suggestion modulates cognitive conflict. *Psychological Science*, 17, 721–7.

Jamieson, G. A. and Sheehan, P. W. (2002). A critical evaluation of the relationship between sustained attentional abilities and hypnotic susceptibility. *Contemporary Hypnosis*, 19, 62–74.

Kaiser, J., Barker, R., Haenschel, C., Baldeweg, T., and Gruzelier, J. H. (1997). Hypnosis and event-related potential correlates of error-processing in a Stroop type paradigm: A test of the frontal hypothesis. *International Journal of Psychophysiology*, 27, 215–22.

Kallio, S., Revonsuo, A., Hämäläinen, H., Markela, J., and Gruzelier, J. (2001). Anterior brain functions and hypnosis: A test of the frontal hypothesis. *International Journal of Clinical and Experimental Hypnosis*, 49, 95–108.

Kinnunen, T., Zamansky, H. S., and Block, M. L. (1994). Is the hypnotized subject lying? *Journal of Abnormal Psychology*, 103, 184–91.

Kirsch, I. (1985). Response expectancy as a determinant of experience and behaviour. *American Psychologist*, 40, 1189–202.

Kirsch, I., Silva, C. E., Carone, J. E., Johnston, J. D., and Simon, B. (1989). The surreptitious observation design: An experimental paradigm for distinguishing artifact from essence in hypnosis. *Journal of Abnormal Psychology*, 98, 132–6.

Kirsch, I., Montgomery, G., and Sapirstein, G. (1995). Hypnosis as an adjunct to cognitive-behavioral psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 63, 214–20.

Kirsch, I., Roberts, K., Mazzoni, G, Dienes, Z., Hallquist, M. N., Williams, J., and Lyn, S. J. (2008). Slipping into trance. *Contemporary Hypnosis*, 25, 202–9.

Kirsch, I. Cardeña, E., Derbyshire, S., *et al.* (2011). Definitions of hypnosis and hypnotizability and their relation to suggestion and suggestibility: A consensus statement. *Contemporary Hypnosis & Integrative Therapy*, 28(2), 107–11.

Kosslyn, S. M. and Thompson, W. L., Constantini-Ferrando, M. F., Alpert, N. M., and Spiegel, D. (2000). Hypnotic visual illusion alters colour processing in the brain. *American Journal of Psychiatry*, 157, 1279–84.

Lau, H. C. and Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Science*, 103, 18763–8.

Lynn, S. J., Kirsch, I., and Hallquist, M. N. (2008). Social cognitive theories of hypnosis. In M. Nash and A. Barnier (Eds.) *The Oxford Handbook of Hypnosis: Theory, Research, and Practice*, pp. 111–140. Oxford: Oxford University Press.

McConkey, K. M., Bryant, R. A., Bibb, B. C., and Kihlstrom, J. F. (1991). Trance logic in hypnosis and imagination. *Journal of Abnormal Psychology*, 100, 464–72.

McConkey, K. M., Szeps, A., and Barnier, A. (2001). Indexing the experience of sex change in hypnosis and imagination. *International Journal of Clinical and Experimental Hypnosis*, 49, 123–38.

McGeown, W. J., Venneri, A., Kirsch, I, Nocetti, L., Roberts, K., Foan, L., and Mazzoni, G. (in press). Suggested visual hallucination without hypnosis enhances activity in visual areas of the brain. *Consciousness and Cognition*.

Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self*. New York: Basic Books.

Miller, M. E. and Bowers, K. S. (1986). Hypnotic analgesia and stress inoculation in the reduction of pain. *Journal of Abnormal Psychology*, 95, 6–14.

Miller, M. E. and Bowers, K. S. (1992). Hypnotic analgesia: Dissociated experience or dissociated control? *Journal of Abnormal Psychology*, 102, 29–38.

Milling, L. S., Kirsch, I., Meunier, S. A., and Levine, M. R. (2002). Hypnotic analgesia and stress innoculation training: Individual and combined effects in analog treatment of experimental pain. *Cognitive Therapy and Research*, 26, 355–71.

Nadon, R., Hoyt, R. P., Register, P. A., and Kihlstron, J. F. (1991). Absorption and hypnotizability: context effects reexamined. *Journal of Personality and Social Psychology*, 60, 144–53.

Nash, M. and Barnier, A. (Eds.) (2008). *The Oxford Handbook of Hypnosis: Theory, Research, and Practice*. Oxford: Oxford University Press.

Oakley, D. A. (2008). Hypnosis, trance, and suggestion: Evidence from neuroimaging. In M. Nash and A. Barnier (Eds.) *The Oxford Handbook of Hypnosis: Theory, Research, and Practice*, pp. 365–392. Oxford: Oxford University Press.

Pettinati, H. (1982). Measuring hypnotisability in psychotic patients. *International Journal of Clinical and Experimental Hypnosis*, 30, 404–16.

Proust, J. (forthcoming). *Philosophy of Metacognition: mental agency and self-awareness*. Oxford: Oxford University Press.

Raz, A. and Campbell, N. K. J. (2011). Can suggestion obviate reading? Supplementing primary Stroop evidence with exploratory negative priming analyses. *Consciousness and Cognition*, 20, 312–32.

Raz, A., Shapiro, T., Fan, J., and Posner, M. I. (2002). Hypnotic suggestion and the modulation of Stroop interference. *Archives of General Psychiatry*, 59, 1155–61.

Raz A., Landzberg K. S., Schweizer H. R., *et al.* (2003). Posthypnotic suggestion and the modulation of Stroop interference under cycloplegia. *Consciousness and Cognition*, 12, 332–46.

Raz, A., Kirsch, I., Pollard, J., and Nitkin-Kaner, Y. (2006). Suggestion reduces the Stroop effect. *Psychological Science* 17(2), 91–5.

Roche, S. and McConkey, K. M. (1990). Absorption: Nature, assessment, and correlates. *Journal of Personality and Social Psychology*, 59, 91–101.

Rosenthal, D. (2005). *Consciousness and mind*. Oxford: Oxford University Press.

Rosenthal, D. (2008). Consciousness and its function. *Neuropsychologia*, 46, 829–40.

Rounis, E., Maniscalco, B., Rothwell, J., Passingham, R. E., and Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness *Cognitive Neuroscience*, 1, 165–75.

Rubichi, S., Ricci, F., Padovani, R., and Scaglietti, L. (2005). Hypnotic susceptibility, baseline attentional functioning, and the Stroop task. *Consciousness and Cognition*, 14, 296–303.

Shea, N. (in press). Reward prediction error signals are meta-representational. *Nous*.

Sheehan, P. W., Donovan, P., and MacLeod, C. M. (1988). Strategy manipulation and Stroop effect in hypnosis. *Journal of Abnormal Psychology*, 97, 455–60.

Spanos, N. (1986). Hypnotic behaviour: a social–psychological interpretation of amnesia, analgesia, and 'trance logic.' *Behavioural and Brain Sciences*, 9, 449–502.

Spanos, N. P., Menary, E., Brett, P. J., Cross, W., and Ahmed, Q. (1987). Failure of posthypnotic responding to occur outside the experimental setting. *Journal of Abnormal Psychology*, 96, 52–7.

Van Nuys, D. (1973). Meditation, attention, and hypnotic susceptibility: A correlational study. *International Journal of Clinical and Experimental Hypnosis*, 21, 59–69.

Varga, K., Németh, Z., and Szekely, A. (2011). Lack of correlation between hypnotic susceptibility and various components of attention. *Consciousness and Cognition*, 20, 1872–81.

Ward, N. S., Oakley, D. A., Frackowiak, R. S. J., and Halligan, P. W. (2003). Differential brain activations between intentionally simulated and subjectively experienced paralysis. *Cognitive Neuropsychiatry*, 8, 295–312.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review*, 101, 34–52.

Woody, E., Z. and Sadler, P. (2008). Dissociation theories of hypnosis. In Nash, M., and Barnier, A. (Eds.) *The Oxford Handbook of Hypnosis: Theory, Research, and Practice*, pp. 81–110. Oxford University Press.

Zamansky, H. S. and Clark, L. E. (1986). Cognitive competition and hypnotic behaviour: Whither absorption? *International Journal of Clinical and Experimental Hypnosis*, 34, 205–14.

Chapter 17

What metarepresentation is for

Tillmann Vierkant

It is widely accepted that many creatures do have minds, but most people agree as well that not many creatures, if any but humans, know that they have minds. Humans not only have beliefs, desires, etc., but they also think about these states all the time. This ability to direct the mind away from the environment and onto itself is not just any old ability either. Many people believe that it is crucially important in explaining why human minds seem to be so much more powerful than the ones of any other creature we know of.[1] It has, for example, been argued that it allows humans to take their minds offline, to break the direct link between perception and action (see, e.g. Prinz (2003) dual representation). Knowing that the bear you contemplate is no more than a thought and not the real thing allows you to think in peace about what would be the best thing to do if you really were to meet one. If you are not able to distinguish between thoughts and the real thing, then the only thing that thinking about a bear will prompt is a fast escape up the next tree. In addition, recognizing thoughts as thoughts allows you to understand that they can be false, and this knowledge should allow you to make much better predictions about your fellow beings. As well, in the case of yourself, it will allow you to not take at face value what you believe now, but to re-examine the steps that led you to a certain belief, and to generally regiment the way you do your thinking.

Thus, it is clear that thinking about thinking is a crucially important ability of humans. What is a lot less clear is what it actually consists of, and whether there really is only one ability that could be referred to as thinking about thinking. Indeed, if one looks more closely at the literature, one realizes very quickly that this is very doubtful. In this chapter, I set out to contrast two ways of what it could mean that humans know that they have minds. In particular, I want to focus on the question of whether we need the ability to metarepresent in order to think about our minds. It will turn out that, at least in some senses of thinking about thinking, this is not a requirement. This might seem surprising, as metarepresentation is supposed to be nothing else than representing representations as such. Indeed, one of the most (if not the most) basic characteristics of the mind is that it represents.[2] So, if you have no concept of what a representation is, how can you think about things that are, at heart, representational things?

The trick that makes it nevertheless possible to think about your thoughts without understanding the nature of representation is that the representations in human minds are not only mental states, but also have contents. Many people have argued that it is possible to think about contents as contents without understanding the nature of mental representational states.[3] Here I want to

[1] Really, all the way back via Strawson to Kant.

[2] Or in less cog sci language, have aboutness or intentionality.

[3] One way to achieve this would obviously be introspection. Many people were very fond of this route, but obviously, there are as well many famous problems associated with it. Ultimately, I think that

look at the attempt to use language as the tool to make this odd sounding claim plausible.[4] Specifically, I will start by looking at the idea of the 'self-regulating mind' as defended by Pettit and McGeer (2002) and explore with them the idea that one way of explaining the difference between human and other minds is that, in contrast to other minds, human minds are self-regulating and that this self-regulation is achieved by thinking about content as content in lingual form. However, Pettit and McGeer do not clearly distinguish between different forms of self-regulation. It is here where this chapter departs from their account and explores which forms of self-regulation do require the ability to metarepresent.

The self-regulating mind

Pettit and McGeer claim in their article 'The self-regulating mind' that what sets human minds apart from the minds of other creatures is that they are self-regulating and not merely routinized. They claim that the crucial difference between the two is that self-regulating minds can do things intentionally to control their beliefs, which routinized minds can't. Self-regulating minds do this by using the unique properties of language. Humans, unlike any other creature, represent the world not only in their beliefs, but also in the sentences that they use to express those beliefs.[5]

It is the properties of this second representational system that, according to Pettit/McGeer make human minds special. On the one hand, language is special because it is very closely related to beliefs. If someone sincerely asserts *p* then she thereby expresses her belief that *p*.

But, on the other hand, language does have some special properties as well which sets it apart from beliefs. The first property that Pettit/McGeer discuss is the fact that language allows for what they call 'content attention'. Routinized minds obviously have contents as well, but as Pettit/McGeer write, they are blind to them. They, as it were, look through them into the world. Language, on the other hand, solidifies the content into what Andy Clark calls 'material symbols' (2006).[6] These material symbols provide a new set of representational objects for the mind to work with.[7] Pettit/McGeer write that this enables attention to contents as such. But it seems more adequate to say that it is not only content that language delivers. Language delivers as well a completely new set of vehicles that bear the contents. Obviously, they only do so in a derived way, but they are content-bearing vehicles nevertheless. The huge advantage of these vehicles, in contrast to the vehicles of beliefs, is that they are visible in the world (in the case of the spoken word this will obviously be replaced by audible) and manipulable for the cognizer. This allows the cognizer

introspection is a dead end, but it is not relevant to defend this view here in depth. See, for example Eric Schwitzgebel's (2008) view for a contemporary critique of introspection.

[4] People who have tried to use language for that claim are, e.g. Dennett (1998) or Clark (2006).

[5] This is not supposed to be a simple on/off thing. In fact, Clark (2006) discusses cases from the animal literature where the presence of material symbols enables chimps to significantly enhance their abilities. Nevertheless, it is obviously equally true that no other animal is anywhere near the same ball park as humans when it comes to mastery of language.

[6] Another important name here is clearly Dennett (e.g. 1998), who developed a similar account to Clark, but sees language as even more transformative.

[7] So really what language does is give the cognizer a second representational system. Once we understand this as what language does, it becomes clear that there could be other ways to achieve this second and/or offline representational stream. Even though I find the idea that language provides this very plausible, I do want to remain neutral about alternative possibilities. What really matters for me is to clarify the relationship between this second representational stream (however constructed), the mastery of folk psychology (or mental state concepts), and the intentional control of the mental.

to direct her attention to all the different properties of the content stored in these visible/audible vehicles. As Pettit/McGeer point out, one of the most important consequences of this is that the cognizer now can learn much more easily about the properties that constrain propositions and thereby learn to build up a network of propositions that respect the identified constraints.

The difference that is most important for Pettit/McGeer, however, is the fact that in contrast to believing, asserting is something that we can do voluntarily. Whether or not someone believes that *p* is normally not up to them. It depends on their evaluation of the evidence for *p*, but whether or not somebody asserts that *p* is entirely under their voluntary control. Obviously, what exactly voluntary control is, is contentious, but here it is enough to say that assertion, but not belief, is voluntary in a very similar sense to lifting one's arm. If the agent decides to do it, then it will normally happen. Pettit/McGeer believe that this difference in intentional control is at the heart of the difference between routinized minds and self-regulating minds. It allows humans to voluntarily implement constraints that improve their thinking. They can now voluntarily give themselves greater exposure to relevant evidence for their deliberations, they can intentionally control their reasoning process, by, for example, repeating steps, reminding themselves of important considerations, and so on. In addition, intentional control even allows humans to overcome very strong natural evaluative tendencies. Pettit/McGeer discuss the example of the pilot who can learn to overcome the tendency to form a belief because of proprioceptive signals and to intentionally use the instrument panel to guide her behaviour, even if it recommends actions that are in sharp contrast to proprioceptive demands.[8]

These are clearly all very important abilities and it seems fair to say that they very plausibly could justify the distinction between non-human routinized minds and human self-regulating minds.[9] However, there is one move in the paper which demands closer attention. In the first half of the paper Pettit/McGeer describe how language enables attention to content and the identification of constraints for this content. But then they prepare the next step in their argument (which is about how humans can intentionally implement content) by saying that humans ascribe beliefs and desires to themselves and that they understand themselves as intentional systems (2002, p. 288). This clearly seems to be about ascribing mental states with specific contents to oneself, rather than ascribing pure content to oneself (it seems incomprehensible, at least to me, to ascribe pure belief contents, but not states). It is one thing to attend to material symbols without attending to psychological states—in fact, this is, as discussed, the beauty of language, i.e. that it does give a new set of content-bearing vehicles, about which we can think without having to have a clue about how these things might work in the human mind. But it is quite another thing to ascribe beliefs to oneself, without understanding that one thereby ascribes a psychological state.

So perhaps understanding mental states is actually far more important for self-regulating minds than Pettit/McGeer at first seem to make out. That Pettit/McGeer might think so is borne out as well by the way their paper progresses. They discuss the importance of folk psychology in many contexts, and again it seems obvious that folk psychology is about ascribing mental states, rather than about attending to and manipulating pure content.

But more importantly, it is not only the fact that Pettit/McGeer seem to think that the ascription of mental states plays a crucial role for the self-regulating mind, which seems to suggest that understanding mental states as states is important for self-regulation. Independent of this ad

[8] This fits very nicely with the mental muscle view of the will as defended by Baumeister (2008) and Holton (2009).

[9] This is obviously not to say that there might not be many other equally plausible ways of accounting for the differences.

hominem argument, it seems right to say that the basic structure of intentional control of minds requires an understanding of mental states as states.

This is because in order to intentionally manipulate anything, it seems necessary that one has some kind of hold on what it is that one is manipulating and in the case of manipulating the mind, that is clearly a mental state.[10] If it were the case, however, that all intentional control of the mental as described by Pettit/McGeer requires metarepresentation, then we would now have found what we really need metarepresentation for, and it would be very important for a very large chunk of human thinking indeed.

But do we really need to be able to ascribe folk psychological states to ourselves in order to intentionally manipulate our mental lives? A first reason to be doubtful about this comes from animal research. In Michael Beran's lab (Evans and Beran 2007), for example, they could show that chimps can use intentional strategies in order to master delayed gratification tasks. When offered the choice between a small reward now or a bigger reward later, the chimps started playing with a toy in order to distract themselves from the urge to take the smaller reward now.[11] This looks very much like the chimps used intentional behaviour in order to prevent a re-evaluation of their belief that the benefits of a large reward later are bigger than the ones of a small reward now, which would have happened if they had focused on the small reward. It actually looks remarkably similar to the pilot case described by Pettit/McGeer. Crucially, however, nobody seems to suspect that the chimps have mastered metarepresentation or possess a folk psychology and were ascribing mental states to themselves. So clearly, the chimps are acting intentionally to bring about a desired result and get the bigger reward, but they are clearly not aware of the mechanism that makes it the case that the strategy that they employ is successful.[12]

But where one might still have some doubts in the chimp case as to whether these animals do not have something like a folk psychology, there are much more banal cases where intentional action prevents or allows the re-evaluation of a given situation. Take, for example, the turning of its head by a nervous mouse in order to see whether the buzzard it was watching fearfully earlier on is still on the tree at a safe distance where it had been before. Turning its head clearly allows the mouse to update, or, if you prefer, to regulate its beliefs, but it would seem ludicrous to suggest that it does so because it is aware that its belief about the location of the predator is by now possibly false.

This is the solution to our puzzle then: it is quite possible to manipulate mental states by acting intentionally, even if you don't know what mental states are, if the manipulation of mental states is not what you intend to achieve but is connected to (or is, for example, the cause of) you successfully reaching a first order goal.

Taking this insight on board, we can now return to the human case and ask ourselves whether we really need metarepresentation in order to go to the library or in order to rehearse an argument? Do we really need to understand ourselves as psychological creatures in order to do these things? The answer is that no, we do not: it seems quite enough if we understand that propositions

[10] Pamela Hieronymi (2009), for example, calls all intentional mental action managerial control and describes this form of control as intentionally manipulating mental objects. She also refers to managerial control as attitude directed control, which really says it all.

[11] The toy was under normal circumstances considered boring by the chimps, so they really only played with it when it was used for distraction purposes.

[12] If chimps can solve tasks like this then one might wonder whether or not pilots will use metarepresentation in order to get their behaviour right. (Thanks to Joelle Proust for pointing this out to me.) I am very happy to concede that many pilots will not use explicit psychological knowledge. Crucial is, that as human metarepresenters they could and it will help them to find successful strategies for achieving their goal.

can be true or false, and that we can find evidence for their truth or falsity by going to the library or rehearsing an argument. Obviously understanding the truth of a proposition is quite different from being able to employ a distracting strategy, as in the case of the chimps, or from the even more basic head-turning behaviour of the mouse. It is in one sense not first order, but thinking about thinking, because, thanks to language, humans but not chimps[13] or mice understand something about the features of content (that it, for example, can be right or wrong, well supported by the evidence, etc.). In a different sense of thinking about thinking, however, it is just as much first order as in the case of the chimps or mice. Going to the library to find better evidence for the proposition you are interested in does not require you to understand that this proposition is the content of a psychological state. You do not have to be able to ascribe such states to yourself and you do not have to be a folk psychology user. Instead, you can simply concentrate on the properties of the content, which are visible to you thanks to their embeddedness in the material symbols of the language vehicle.[14]

So we actually do not need to be able to employ the intentional stance[15] in order to intentionally regulate our minds, but obviously, as long as we have not mastered mental state concepts we will not be aware of the fact that that is what we are doing. At the same time, it could still be true that it is thinking about thinking that makes self-regulation in the human case categorically different from the intentional control that mice and chimps have over their minds. Language gives humans a huge range of new objects that they can intentionally manipulate (think about) which are not available to animals. These objects that human intentions can be about will be the propositions we have in the form of lingual material symbols. So even though I disagree about the importance of folk psychology for being able to use language as a self-regulation tool, it might obviously still be the case that the intentional control which is enabled by language is what makes all the difference between us and animals. On the other hand, it might not be. It might be as well the other features of language discussed here that make all the difference.[16] Whatever the right answer, what matters in this context is that there are intentional forms of self-regulation which do not require that the agent understand anything about mental states. This leaves us now with the question of whether this understanding is simply irrelevant in the context of the intentional regulation of the mind. It is this question to which we will now turn.

Mindreading

In the previous section I argued that folk psychology is not necessary for intentionally regulating one's mind. This ability may, combined with language, be crucial for explaining the difference

[13] This is probably not an on/off phenomenon and it is therefore very likely that chimps do already possess a few material symbols. See, e.g. Clark (2006).

[14] In this respect, Hieronymi seems to have got it wrong, if she thinks that managerial (i.e. intentional) control of mental states must be attitude directed—at least if attitude directed means that the attitude is what the agent is aware of targeting in the action.

[15] We have to be careful here, because it is not clear whether the intentional stance as developed by Dennett (1987) really is a psychological stance. If we give it a behaviourist reading, then it might actually be reducible to the kind of thing that language enables. On such a reading, however, the intentional stance is then unsuitable to understand humans as psychological creatures.

[16] I take it that Pettit/Mc Geer would probably be on one side here, Mele (2009) and Holton (2009) somewhere in the middle, and Moran—even though I do think there is some ambiguity in the notion of deliberation as employed by Moran (2001)—and Hieronymi (where it sometimes feels as if deliberation actually does include non aware intentional control), and Strawson (2003) on the other.

between animal and human minds, but attributing mental states to oneself is not necessarily part of this form of self-regulation.

From this conclusion we now move naturally into the second part of the paper: if self-regulation in the sense described in the last section does not require metarepresentation, what do we need it for then? Which are the situations where it does matter whether we know that the material symbols we are manipulating derive their intentionality from the beliefs of people? The obvious candidate here is the ascription of false beliefs. As long as we do not understand that people's beliefs can deviate from the rational norm, it seems very difficult to imagine how one can predict that someone will act on or even that someone simply has a false belief. Obviously, this was exactly the reasoning that led to the debate defining false belief task (Wimmer and Perner 1983).

Nowadays, however, there are some significant worries about the need for a folk psychological theory in solving false belief tasks. The problem is that false belief task-like competencies have been found not only in preschoolers, but also in toddlers (e.g. Buttelmann 2009) and infants (e.g. Onishi 2005). If we do not want to claim that these (in the case of the infants: prelingual!) children possess mental state concepts, then it must be possible to solve false belief tasks without such concepts. But if it is possible to do that, then there is now a heavy burden of proof on defenders of the theory-theory, who claim that children use theory in the later explicit tasks (see, for example, Perner (in press) for Povinelli and Voh's challenge and Perner's reply to it).

But whether or not this challenge can be successfully answered, there is an additional reason to be sceptical about the importance of possessing mental state concepts for our discussion here. This is because our topic is not mindreading, but the function of metarepresentation in metacognition. In other words, our question is: what is the function of theory-based mindreading for self? But if it is the case that mindreading is important in the case of other minds, because others might not share our beliefs and we need the psychological theory to help us to understand this psychological fact so that we can make better predictions, then it seems clear that the case of our own minds will be quite different. After all, as Moore (1942) has famously pointed out, it does sound very odd indeed, if one asserts a proposition, but denies that one believes it.[17]

What is the use of mindreading for self?

One of the most discussed worries for theory-theory always was that it does not respect the intuitively obvious difference between self-knowledge and knowledge of other minds.[18] It is intuitively quite plausible that I might need a piece of theory in order to figure out what you are thinking, but it does seem very counterintuitive that the same should be true when I want to know what I am thinking.[19] Obviously, the literature within philosophy that tries to explain what it is that makes knowledge of our own minds special is vast, but we will be only interested in the explanation that has been derived by Pettit/McGeer directly from the idea of the self-regulating mind. According to the so-called agency theory of self-knowledge, what is special about self-knowledge is that, in order to find out what you believe, you do not have to do any psychology, but simply deliberate about the first-order question. Once you have found your answer to the first-order question, you have as well answered your belief question. One has special authority about one's own beliefs, because by answering the question, one creates (or at least makes visible) the

[17] Moore famously claimed that to assert that 'I went to the pictures last Tuesday, but I don't believe it' seems absurd. This is referred to as Moore's paradox in philosophy. They are particularly interesting because they seem paradoxical even though there is nothing obviously logically wrong with them.

[18] For a very good introduction to the debate between theory theory and simulation theory see Davies (1995).

[19] Editors' note: for a dissenting opinion, see Carruthers and Ritchie (Chapter 5, this volume).

relevant state.[20] Importantly, in this context, if self-knowledge is created because it automatically flows from the deliberations of the agent, then it cannot normally go wrong. It is quite different from normal observational knowledge and therefore it seems difficult to see why we would need any psychological theory to improve it.

Self as other

The rule that self-knowledge is gained by deliberating about first-order questions has exceptions, however—when we are interested in our psychology not because of the contents of its states, but because of their nature as states. This interest could exist for two quite different reasons. On the one hand, we might think that certain mental states have a positive (or negative) value as states rather than as contents, while on the other hand, it might be the case that we are interested in mental states as states, because this allows us to see ourselves as psychological beings rather than as rational agents. This allows us to take into account the fact that our psychology might change and we might one day think something is rational which we now would consider plainly irrational, or that there might be psychological traits in us that influence our thinking and acting even though they do not appear in our rational justifications.

A very nice example of the former is Pascal's wager. Pascal felt that he had a very good reason to believe in God, even if his existence was very unlikely. He argued that the belief would not do much harm, even if it were wrong, but that the consequences would be dire if one did not believe and it turned out that God existed. One problem for Pascal was that it is not clear how knowing that it would be good to have the belief in God would help in any way in acquiring it. Pascal's solution to this problem is ingenious. He argued that it is quite possible to acquire the belief, even against one's rational evaluative tendencies, by simply conditioning oneself in the right way. So Pascal recommended going to mass and praying regularly and over time the sheer force of habit would produce the belief that seemed out of reach by rational means. Pascal uses psychological knowledge in order to achieve the odd state of not believing a proposition (he believes that God probably does not exist) and at the same time asserting confidently that he will believe it very soon (because he knows he will have conditioned himself to do so).

A similar case is Kavka's famous toxin puzzle. In it, an agent gets offered a large sum of money, if at midnight she has the intention to drink a mildly unpleasant toxin the next day. The puzzle arises, because it seems very difficult for a rational agent to collect the reward. This is because as the agent gets nothing for actually drinking the toxin, but only for having the intention, she has absolutely no reason to drink it when the time comes. As she knows this, it becomes impossible for her to *intend* to drink it, because in order to intend something the agent has to be settled on doing it, but the agent knows that there is no reason for actually drinking it. As puzzling as this story is, it obviously only works as long as the agent is not allowed to form the intention by non-rational means.

Future directed self-control

In his book *Reasons and Persons*, Derek Parfit (1984) includes the example of a rich, young communist who ardently believes in material equality, but who also knows that statistically most rich, young communists turn into rich, old conservatives. The communist faces the dilemma of

[20] Pettit/McGeer build here on the work of Richard Moran. However, it has to be mentioned that there is an important ambiguity in their usage of Moran's work. Deliberation Moran style is not an intentional affair, whereas in Pettit/McGeer it is at least partly intentional.

whether he should give away the money now, and therefore be true to the ideals he now holds, or keep it, thereby maximizing his chances of doing what he will come to think is the right thing to do and which, given the assumption that he will be a rich conservative for longer than he is a rich communist, is what maximizes doing what he thinks is the right thing over his lifetime. The example is wonderful, because it illustrates how incredibly deep the chasm between our first-order thinking and our psychological perspective on ourselves can be. The rich conservative that he will be is worrying for the communist, not because he thinks that by the time he is old he will be too weak-willed for his ideals, but because he believes the evidence that suggests that older people are more likely to genuinely judge that it is morally wrong to give up the property that has been in the family for generations. He knows that, even though he is connected to his future self by means of a continuous and slowly changing psychology, it is quite likely that the values that make his decisions in an important sense his as an agent on many accounts of free agency will have been replaced by ones he now deplores.[21] Of all the deep issues on ethics and personal identity that the example raises, what matters for us here is that it illustrates incredibly well one important aspect of what our psychological perspective on ourselves adds to first-order reasoning. It allows us to see our future selves as somebody else[22] and to realize that even our deepest-felt convictions are possibly merely transient psychological states.

Most importantly, however, realizing that our first-order evaluations may change enables us to have a completely new level of self-control. As long as the agent does not have psychological knowledge, it will be nigh on impossible for her to conceive that something which she very strongly believes to be true now could be judged by her to be false by tomorrow. As she can sincerely see no evidence that would render the proposition in question false, it becomes very difficult, if not impossible, for her to comprehend that she might nevertheless judge very differently tomorrow.

This, then, is the difference between the intentional control of the mental that is enabled by having a new set of material symbols which one can intentionally manipulate and the control that is dependent on the understanding of the fact that one is a psychological creature. Only the second allows for targeted interventions that aim at the states rather than the contents.

The targeted interventions that one can only use competently and flexibly if one understands oneself as a psychological being are the many old self-control strategies that humans have so successfully invented over the course of their history. They range from providing external constraints on undesirable action options (like tying yourself to the mast to prevent yourself from jumping to your death or giving your car keys to the landlord in order to prevent yourself from achieving the same result by driving), to directly influencing your psychology. These self-control goals can be achieved by providing relevant input for the machinery (remind yourself what a hangover feels like) or even by direct tampering with the machinery (a couple of drinks will make you less shy). Importantly, as we have seen with the chimps, there are only very few things in self-control which you cannot do *at all* if you do not understand mental states (Pascal-like cases are the only ones that spring to mind), but understanding that you are a psychological creature means that you understand the psychological side of self-control and this in turn allows a whole new league of flexibility in employing self-control tools (unlike the chimps, humans know that distraction can help to achieve the bigger reward by preventing changes in beliefs).

[21] In Parfit's architecture, this example serves to show that we should not overestimate the importance of doing what maximizes our utility in our life as we perceive it across time rather than maximizing what we now think would be the right thing to do with our life.

[22] We have to be careful here, because it is by no means certain that we really most of the time use the psychological stance when interpreting others.

Mindshaping

Before we come to the end, it seems very important to put the argument in this chapter in relation to a related but different recent development in the mindreading literature. Based on loosely Dennettian roots, thinkers like Dan Hutto (2008) or Tad Zawidsky (2008) have recently made a strong case for the idea that folk psychology might not be for mindreading, but for mindshaping. The core of the idea is that human development is scaffolded by the constant ascription of propositional attitudes to children, even at an age when these children do not really understand these terms. These attributions act then like self-fulfilling prophecies. They provide a normative standard that the children try to meet.

This seems to me to be a powerful account of the use of folk psychology, but if it is true, one might think that it undermines the position defended here. For one, it seems to show that folk psychology can be used to shape minds, even at a time when children do not yet understand metarepresentation, and secondly it seems to show that theory-based mindreading is not necessary for shaping minds later either, because the mindshaping happens simply because of the practice of ascribing mental states to each other, which then act as normative markers. All of this seems to be possible without the need for theory.

In reply to this, it should be pointed out that this chapter is mainly interested in the *intentional* self-regulation of the human mind. Mindshaping does not really have an awful lot to say about that, because the effects of ascribing folk psychological states are largely automatic (it is not that we try to conform to them, but taking ourselves to have them leads to automatic adjustments). Secondly, the argument about the use of language developed in this chapter is supposed to show that language-related thinking about thinking gives massive cognitive advantages that are quite unrelated to folk psychology. In this respect, then, language-based thinking about thinking is much wider than the scaffolding provided by ascribing folk psychology terms to children. Finally, it is again the issue of flexibility that singles out theory-based self-regulation as discussed in the second half of the chapter. Once the agent is aware of what it is that she is doing, self-regulation can become much more efficient and flexible than in a case where mindshaping happens only in the unreflected normative marker way as described by Hutto or Zawidsky. All in all, then, mindshaping seems highly compatible with the ideas defended here.

Conclusion

In this chapter we examined two different ways in which thinking about thinking enables intentional self-regulation. On the one hand, we looked at the way in which language provides us with a new set of representational objects for the mind to manipulate. The great thing about these objects, in contrast to the vehicles of beliefs, is that they can be intentionally manipulated in a very straightforward way. Agents can choose what to assert and this means that it becomes much easier to attend to (rehearse, refine, etc.) contents. However, it became clear that this ability does not require one to attribute or even understand the nature of mental states. It is not the mastery of folk psychology that enables this control. In the second half of the chapter we then asked ourselves whether there is some form of intentional self-control which does require the mastery of folk psychology. We discussed two forms of control where this seems to be the case. Obviously, an understanding of mental states is required, if what we are interested in is the acquisition of such states as states. However, these cases do seem rather rare, so it was the second form on which we concentrated. We argued that an understanding of mental states is crucial for forms of intentional self-control where we need to take into account the fact that our psychology can change over time. In these cases, as long as we do not have an understanding of ourselves as psychological beings, we will struggle to effectively achieve our self-control aims. Folk psychology might not be

what separates humans from animals, but it seems likely that if, as most people seem to assume, self-control is crucial for autonomy, an understanding of folk psychology is crucial for an autonomous agent.

References

Baumeister, R. (2008). Free will, consciousness and cultural animals. In J. Baer (Ed.) *Are we Free?*, pp. 65–85. Oxford: Oxford University Press.

Buttelmann, D., Carpenter, M., and Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112, 337–42.

Clark, A. (2006). Material symbols. *Philosophical Psychology*, 19(3), 1–17.

Davies, M. and Stone, T. (Eds.) (1995). *Folk psychology*. Oxford: Blackwell.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. (1998). Reflections on language and mind. In P. Carruthers, and J. Boucher (Eds.) *Language and thought: interdisciplinary themes*, pp. 284–94. Cambridge: Cambridge University Press.

Evans, T. A. and Beran, M. J. (2007). Chimpanzees use self-distraction to cope with impulsivity. *Biology Letters*, 3, 599–602.

Hieroymi, P. (2009). Two kinds of mental agency. In L. O'Brien and M. Soteriou (Eds.) *Mental Actions*, pp. 138–62. Oxford: Oxford University Press.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Hutto, D. (2008) *Folk Psychological Narratives. The sociocultural basis of understanding reasons*. Cambridge MA: MIT Press.

Mele, A. (2009) Mental actions a case study. In O'Brien, L. & Soteriou, M. (Eds.) *Mental Actions*, pp. 17–37. Oxford: Oxford University Press.

Moran, R. (2001). *Authority and Estrangement: An Essay on Self-knowledge*. Princeton, NJ: Princeton University Press.

Moore, G. E. (1942). A reply to my critics. In P.A. Schlipp (Ed.) *The Philosophy of G. E. Moore*, pp. 660–7. Evanston, IL: Northwestern University.

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255–8.

Parfit, D. (1984). *Reasons and Persons*. New York: Oxford University Press.

Perner, J. (in press). Theory of mind – an unintelligent design: From behaviour to teleology and perspective. In A.M. Leslie and T. German (Eds.) *Handbook of 'Theory of Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pettit, P. and McGeer, V. (2002). The self-regulating mind. *Language and Communication*, 22, 281–99.

Prinz, W. (2003). Emerging selves. Representational foundations of subjectivity. *Consciousness and Cognition*, 12, 515–28.

Strawson, G. (2003). 'Mental Ballistics Or The Involuntariness of Spontaneity.' Meeting of the Aristotelian Society, University of London, 28 April.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review*, 117, 245–73.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–28.

Zawidzki, T. (2008). The function of folk psychology: mind reading or mind shaping? *Philosophical Explorations*, 11(3), 193–210.

Chapter 18

Anoetic, noetic, and autonoetic metacognition

Janet Metcalfe and Lisa K. Son

Introduction

Metacognition can take many guises. Consider, first, one contestant of several, W., playing a television game show that tests general knowledge by presenting whimsically phrased cues. As a question 'What is Will's Quill?' is displayed on the screen, W. very quickly retrieves bits of information regarding what may possibly be or be not related to the answer, based on the question and his understanding of the natural language associates to it. He accumulates the fragmentary information resulting from his memory search quickly as the clock ticks. If the information count reaches a criterion, but one far less than is necessary for complete access to the answer, W. buzzes in to beat out his opponents and to indicate that he thinks that he *will know* the answer given an additional 5 seconds, even though he does not know it yet. If the accumulation of partial information does not reach criterion, W. declines to respond, letting the opposition buzz in, instead. Using this 'game-show' strategy (Reder 1987) based on the metacognitive feeling that he will know, W. is nearly always—roughly 85% of the time—able to come up with the answer later when he thinks he will be able to do so. And, by combining his encyclopaedic knowledge, his lightning speed, and his sophisticated metacognitive strategy, W. becomes the new world Jeopardy champion.

Now imagine L., who is playing a memory gambling game. He is presented with the target—a complex picture—in a flash on the screen. The picture disappears, and nine alternative pictures appear on the screen simultaneously. L. looks through them considering each in turn and upon seeing what he thinks is the target picture in the array, he touches it, and they all disappear. Then, though, he has to give his confidence in his answer. He can either 'pass'—choose not to wager—or he can 'double down'—wager big. Two betting icons appear on the screen. Nothing further will happen until he makes this retrospective decision about whether he thinks he was right or wrong. In this case, L. chooses the 'double down' icon, and he wins three tokens, which fall into his hopper, to be redeemed later when he has accumulated enough tokens for a prize. Had he pushed the 'pass' icon, he would have gotten only one token. But had he touched the wrong picture in the 9-option task, and then 'doubled down', he would have seen three tokens fly out of his hopper and disappear. L. is known, by other gamblers, as having a *serious* emotional reaction when this happens. But, fortunately, it doesn't happen often. And he does get paid off with prizes from time to time. Like other gamblers, L. is happy to play this game of making metacognitive bets on his own memory hour after hour, day after day.

Finally, imagine S. trying to retrieve the name of the famous Canadian author who wrote *The Last Spike*. A nagging feeling of having the answer right on the tip of his tongue plagues S. But S. cannot retrieve the answer no matter how hard he tries, and he is trying hard. His friends tell him to give up. None of them are Canadians, and they neither know the answer, nor care, to be sure. But S. refuses to listen. His mind is screaming with this impossible-to-resist emotional premonition

that the answer is imminent (see Schwartz and Metcalfe, 2011). And he is right, statistically, at least. When people have this feeling, they nearly always get the answer eventually. But it is hours, not moments away. Having been driven almost to distraction by this tantalizing gap in his knowledge, and knowing that the answer, oddly, is 'almost' a French name, and that the first letter of that first name is P, finally, in a flash of insight the answer—Pierre Berton—appears, seemingly unbidden out of the blue (previous intense efforts to find it notwithstanding).

Which one of the above individuals is metacognitive? Which was making an assessment about an internal representation? Which, by virtue of this metacognitive reflection, has a self? Insofar as all three of these cases represent what many researchers in the field affirm as true metacognition—knowing about what one knows—then, it would seem that the case could be made that all three of them involve these characteristics and each of W., L., and S. exhibit self-awareness. Indeed, a number of distinguished thinkers have forwarded the idea that a central reason for interest in metacognition, above and beyond its functional usefulness in allowing people better control of their thinking and their action, is that metacognition is the key to a special kind of human self-reflective consciousness that is the very essence of our humanness.

Metacognition, by this view, is thought to be what we might call self-perspectival (see Descartes 1637; Husserl 1929; Searle 1992). The emphasis on the relation of metacognition to the self undoubtedly stars the work of Descartes, who reflected about his reflections and perceptions, and in so doing made the claim—that he certainly believed was self-evident and irrefutable—that the fact that he was able to do this reflection provided incontrovertible evidence for the self. 'I think therefore I am' with the 'I' highlighted. The reflection gave the proof of his self. While some moderns, notably Bertrand Russell (1997),[1] are not so sure, it is a fascination with the *self* in self-reflection—that this kind of recursive cognition gives rise to consciousness and self-awareness and proof that an internal person exists—that provides the intellectual glitter giving studies in metacognition their panache.

The modern theorist most associated with this view is Rosenthal (2000). In advancing his 'higher order thought' (HOT) hypothesis, he argues that consciousness is essentially metacognition, which, classically (see Nelson and Narens 1990) entails the reflection at the metalevel upon a lower, basic, level. Rosenthal notes: 'The leading idea behind the HOT hypothesis is that a mental state is conscious only if one is, in some suitable way, conscious *of* that state ... A conscious state is a state one is conscious of oneself as being in' (pp. 231–2). Rosenthal's HOTs involve something more than just a metalevel reflection on a basic level representation: *self*-consciousness is implied. He does not necessarily endorse an elaborate folk-theoretic notion of what self consciousness entails including being explicitly conscious of oneself as the subject, or of having all of one's conscious thoughts and experiences come together mentally. Self-consciousness could be much more pared down: 'HOTs can, instead, represent the self in some minimal way, for example, in terms simply of a distinction between oneself and everything else'. But, even though minimal, some form of self-consciousness is implied. Furthermore, Rosenthal says that such consciousness can only be found in creatures; presumably, computers need not apply. But, perhaps non-human animals could.

[1] Russell notes (p. 17): '"I think, therefore I am" says rather more than is strictly certain. It might seem as though we were quite sure of being the same person to-day as we were yesterday, and this is no doubt true in some sense. But the real Self is as hard to arrive at as the real table and does not seem to have that absolute, convincing certainty that belongs to particular experiences. When I look at my table and see a certain brown colour, what is quite certain at once is not "I am seeing a brown colour", but rather, "a brown colour is being seen". This of course involves something (or somebody) which (or who) sees the brown colour; but it does not of itself involve that more or less permanent person whom we call "I".'

Animal metacognition researchers almost invariably allude to the self-awareness aspect of metacognition in motivating their investigations of whether animals might be able to do metacognitive tasks. For example, Smith et al. (2009) justify their research on animals by saying: 'Metacognition is linked to self-awareness … because doubt is so personal and self-oriented. Metacognition is linked to declarative consciousness, because we can introspect and declare states of knowing. Thus, metacognition is a sophisticated capacity in humans that might be uniquely human' (p. 40). Smith (2009) says 'one of comparative psychology's current goals is to establish whether non-human animals (hereafter, animals) share humans' metacognitive capacity. If they do, it could bear on their consciousness and self-awareness too' (p. 389). Foote and Crystal (2007), who investigated metacognition in rats, say 'People are sometimes aware of their own cognitive processes. Therefore, studies in metacognition test the hypothesis that animals behave functionally the same as an organism that is aware of its own cognitive state' (p. 1).

And, while, if W., L., and S. were all people, we would have no qualms about admitting that the stream and quality of the metacognitive thought processes would allow us to attribute selfhood to each—they 'feel' like people—when we realize that two of these three were not even humans, we might balk at this conclusion. And, indeed, W. in our earlier example, is Watson, the IBM computer who recently made front page news by beating out previous Jeopardy champions to become the new world champion. The feat is impressive, but does it imply that W. is conscious and has a self? And L. is Lashley, a rhesus monkey. S. is human, with the initial chosen for 'Self.' In that light, S.'s musings about his tip-of-the-tongue state leave little doubt, in most people's minds, that he has mind, consciousness, and self-awareness. But while, intuitively, we reject the idea that Watson might have a self, and remain agnostic about Lashley (while perhaps swayed toward the possibility by the metacognitive data), the question remains: if the evidence for self awareness is metacognition, why do we accept that evidence for Self but not for Watson? Perhaps we are merely exhibiting an anthropocentric bias, and the impressive performance on the metacognitive tasks, by all three actors, should mean that we should, rationally, be compelled to abandon our prejudices against machine or monkey and attribute consciousness and a self to all three. One possibility, though, which we explore in this essay, is that perhaps it is only *certain* metacognitive tasks, with *particular* characteristics that imply high-level consciousness and selfhood. We will here endeavour to analyse tasks that have been labelled as 'metacognitive' into three different levels, borrowed from Tulving's (1985) analysis of different levels of consciousness: *anoetic*, *noetic*, and *autonoetic*.

Three levels of consciousness and metacognition

Before analysing various metacognitive tasks we will first review Tulving's (1984; Wheeler et al. 1997; Rosenbaum et al. 2005) distinction between three different levels of consciousness.

Anoetic consciousness

At the lowest level, Tulving defines anoetic consciousness as a state that is temporally and spatially bound to the current time. Although it is a kind of consciousness, it is not one that allows escape in any way from the here and now, and so an animal functioning at this level of consciousness is stimulus bound. A judgement that refers to something in the world even though that something is interpreted through the viewer's perceptual biases and learning would, then, be anoetic. Thus, if a person were learning to discriminate between Pinot Gris and Pinot Grigio, for example, and made judgements, based on tastes of various wine samples, these judgements—being about something in the world, even though the internal percept experienced is, undoubtedly, biased by the learning mind—would be anoetic. Note that while mental processes and past discrimination

learning may interact with just what the subject perceives (we make no claim that perception is naive) the percept, itself, is bound to the moment. It is not a representation or a memory of Pinot Grigio, but rather the percept of the wine itself that is being judged (and so is neither a judgement about an internal representation nor, indeed, is it a judgement about the judgement). By some definitions (see Metcalfe and Kober 2005; Carruthers 2011) a judgement at this level would not be considered metacognitive at all. It would simply be a judgement about the world as perceived. But other researchers (e.g. Reder and Schunn 1996; Smith 2009) have labelled such judgements metacognitive. The framework specified by Nelson and Narens (1990), proposed that there are at least two levels of cognition interacting to form a metacognitive system, a basic level and a meta-level. The basic level, in this anoetic case, would not be a representation at all, however, but rather a percept, and so it is not clear that the word meta-'cognition', should be applied to judgements, such as these, concerned with percepts. They might better be called metaperceptual. But perhaps to overcome the definitional disputes about whether judgements about objects or events in the world as perceived by the subject are metacognitive, and, hopefully to forward our understanding of whether or not self-awareness is involved, we could agree to call such judgements anoetic metacognition. Anoetic consciousness, of course, makes no reference to the self. Similarly, anoetic metacognition could not be considered to involve self-awareness.

Noetic consciousness

This kind of consciousness involves internal representations, and is associated with semantic memory. It allows an organism to be aware of, and to cognitively operate on, objects and events, as well as relations among objects and events, in the absence of the physical presence of those objects and events. Noetic metacognition would be a judgement that is made about a representation. The object on which the judgement is made has to be mental and internal rather than physically present, to qualify as being noetic rather than anoetic. To our knowledge, all researchers agree to call such judgements about mental representations metacognition. However, noetic consciousness, while a form of consciousness as the name implies, does not necessarily involve the self or anything self-referential.

Autonoetic consciousness

This is the highest form of consciousness and is self-reflective or self-knowing. For the first time, the self, then, is intimately involved. This level of consciousness is often, in Tulving's framework, related to human adult episodic memory, which may involve mental time travel of the self. Autonoetic consciousness is thought to be necessary for the remembering of personally experienced events, as long as the memory of those events is self-referential. An individual could not remember something that they experienced in a noetic manner, if they did not know that they had explicitly experienced it, as has been shown to be the case with certain amnesic patients, such as K.C., who are thought to lack autonoetic memory (Rosenbaum et al. 2005). But when a normal person remembers an event in which they participated, he or she is normally thought to be aware of the event as a veridical (or sometimes non-veridical) part of his own past existence, and the involvement of the self is a necessary component in this kind of consciousness. Autonoetic consciousness is not mere depersonalized knowledge. Rather, as James (1890) says: 'this central part of the Self is *felt* ... and no *mere* summation of memories or *mere* sound of a word in our ears. It is something with which we also have direct sensible acquaintance, and which is as fully present at any moment of consciousness in which it *is* present, as in a whole lifetime of such moments' (p. 299). A normal healthy person who possesses autonoetic consciousness is capable of becoming aware of her own projected future as well as her own past; she is capable of mental time travel,

roaming at will over what has happened as readily as over what might happen, independently of physical laws that govern the universe. According to Tulving (2005) only humans past infancy possess autonoetic consciousness.

Do any kind of metacognitive judgements necessarily involve autonoetic consciousness? It would seem that if the judgement makes specific reference to the self it would qualify. A metacognition at the autonoetic level might also be a judgement about one's own personal memories of one's own personal past. From the standpoint of relating metacognition to self-awareness, then, these particular kinds of metacognitions, if there are any such, are of particular importance, since it is only these that involve self-consciousness.

In the sections that follow we will sort metacognitive tasks that have been conducted, both in humans and in animals, into anoetic, noetic, and autonoetic metacognition, with the view to clarifying the use of this reflective (but perhaps not *self*-reflective) processing as a litmus test for ascertaining whether or not particular creatures and, indeed, sophisticated machines, might have self-awareness.

Anoetic metacognition: stimulus-driven judgements

The lowest level of metacognition is anoetic. Any judgement where the individual is evaluating an external stimulus is here categorized as anoetic. Consider the simple example when judging the value of an item, say, a mug. One could say that a mug is worth £10. One's judgement of the mug changes, though, depending on who owns the mug (Kahneman et al. 1990). While the object is 'endowed' with higher value when possessed by the individual (Thaler 1980), as given by his or her subjective judgement, the judgement is, nevertheless, of an external stimulus rather than a representation; it is anoetic and no self-awareness is involved. The judgement of the Pinot Grigio mentioned earlier, whether by a trained or untrained palate, also falls into this category, as do all such perceptual/categorical judgements.

While Foote and Crystal (2007) have argued that rats are able to reflect on their own mental processes, their task was anoetic. The experimenters had their rats learn by reinforcement to discriminate between the duration of two-tone classes. Then they combined this task with one in which the animals, before making the discrimination choice, could pick one response if they wanted their upcoming discrimination choice to let the response count and another (a 'pass' response) if they did not. When the stimulus duration was in the middle of the two learned classes some, but not all, of the rats chose the 'pass' response. Although arguments have been made that the entire sequence was simply a complex chain of conditioned responses (Staddon et al. 2007), even if we allowed that the rats really made a choice to take the test or not, the task is nevertheless anoetic. It was about a categorization of a stimulus in the world, not a representation, and was, in no way, self-relevant.

Similarly, the classic 'escape' studies in dolphins are anoetic. In one such study (Smith et al. 1995), dolphins were required to discriminate the auditory frequencies of two tones by responding with one of two responses. If a 2100-Hz tone was sounded, the dolphin was rewarded when it responded to a '2100-Hz' icon; for all lower frequencies, the dolphin was rewarded when it responded to a '<2100-Hz' icon. An error terminated the trial without reinforcement and resulted in a punishment in the form of a time out. A response to a third 'escape' icon also terminated the trial, but with neither reward nor punishment. It simply acted as an expression of 'I'd rather opt out of this question' and moved onto a new trial. Dolphins could do this task, and sometimes chose to escape rather than take the test. Even allowing that their doing so was a judgement, it was an anoetic judgement, and hence does not imply self-awareness. Other 'escape' type studies (e.g. Smith et al. 1998; Shields et al. 2005; Washburn et al. 2010), where the probe or percept, and not an internal representation, gives rise to the judgement, would also be included as examples of

anoetic metacognition (see Terrace and Son (2009), for a review of yet other cases of anoetic metacognition using the escape paradigm).

It is possible, of course, that monkeys, dolphins, and even rats, have self awareness. But none of the tasks outlined in this section require it. Even those tasks that require a human to simply make a judgement about the world is not evidence that people are self aware (indeed, it can be argued that such a confidence judgement is not metacognitive, but simply, a memory judgement). In the case of humans, however, any judgement that is categorized as 'anoetic' might include self awareness—and thus, be truly metacognitive—given that we can make further judgements about our judgements, verbally. Non-verbal animals are not as fortunate. Even if we agree that anoetic metacognition *is* metacognition—a proposition that we might consider to be stretching the definition of metacognition to the breaking point—it is still anoetic, and does not imply anything about whether or not the organism showing such a capability has a self, or can reflect upon that self in any way.

Noetic metacognition: judgement about an internal representation

Noetic consciousness allows an organism to be aware of, and to cognitively operate on, objects and events, and relations among objects and events, in the absence of those objects and events. The main difference between noetic metacognition and anoetic metacognition is that with the former the judgement is made about an internal representation that is no longer present in space and time, rather than about a stimulus that is present in the world.

Classic cue-only delayed judgements of learning are a typical case of noetic metacognitive judgements. A learning event, consisting of a cue and a to-be-learned target, is presented, and then at some later time, the person is given the cue and asked to make a judgement about whether he or she will later be able to give evidence that they know the target. If they think they will know it, they give it high judgement; if not then they give a low judgement of learning. Note, if people mentally projected their selves into the future to see whether they would get the answer this judgement would be considered autonoetic. However, the data on what people actually do to make this assessment suggest that they do not so mentally time travel. The most compelling evidence for a lack of mental future projection is that people's judgements of learning do not distinguish between whether the test will be 5 minutes or 1 year hence (Koriat et al. 2004)—a distinction that would be large were people really mentally projecting into the future. What they appear to do instead (Son and Metcalfe 2005; Metcalfe and Finn 2010) is first try to recognize the cue. If they cannot do so, they say that they don't know and give a fast low rating. If they do recognize it, they then attempt to retrieve the target, with judgements of learning getting lower and lower the longer it takes them to do so. Thus, the judgement is about the current retrievability of the cue and target, and hence noetic in nature.

Another case of what is probably a noetic metacognitive judgement occurs in the hindsight bias paradigm. After a person has made an assessment about some event and is then given feedback concerning the correct answer, they are asked to remember what their earlier judgement was. They tend to think that their earlier judgement was much closer to the correct answer, which they now know, than it really was (Hoffrage and Pohl 2003). This reflects a hindsight bias or a 'knew it all along' effect. Hawkins and Hastie (1990) defined hindsight as 'a projection of new knowledge into the past accompanied by a *denial that the outcome information has influenced judgement*' (p. 311). In contrast to this idea, though, it seems plausible that the hindsight bias results from a *lack of* projection of the self back into its past state of knowing. The failure to do the past projection, itself, results in the bias. If so, then the judgement is noetic: based, not on mental time travel but rather on current knowledge.

While many experiments indicate that animals have anoetic metacognition, examples of noetic metacognition in animals are much rarer. There are two cases, however, that qualify. In a sequence of trials, Hampton's (2001) monkeys were shown a target picture to study. Then, after a short delay (which was important because it meant that the monkey had to rely on a representation rather than a stimulus currently present in the world), they saw the target picture again, along with three distractor pictures. The monkeys' task was to select the target. However, after seeing the sample and prior to receiving the test, Hampton gave the monkeys the choice of either taking the test, or opting out. On some mandatory trials, though, they had to take the test. The finding of most interest was that the monkeys were more accurate on self-selected test trials than on mandatory trials, suggesting that the monkeys opted out when they knew they did not know the answer. Crucially, they did so when no external stimuli were available as cues at the time of their decision, which means that the judgements were based on internal representation and hence were noetic. However, insofar as no self-reference was necessary, these judgements were not autonoetic.

Finally, Kornell et al. (2007), asked monkeys to make retrospective judgements after they took a memory test. In one such task, monkeys performed a memory task and were then asked to 'wager' on the accuracy of their memories. They first studied six images that were presented sequentially on a touch-sensitive computer screen. Then, one of the six images was presented along with eight distractors and the task was to touch the picture that was already seen in the initial exposure sequence. Once a monkey had touched his choice, he made a wager. Making a 'high' wager meant that he would earn three tokens if his memory response had been correct, and lose three tokens if it had been wrong. Making a 'low' wager meant that he would earn one token, regardless of the accuracy of the memory. Tokens were accumulated at the bottom of the screen and could be exchanged for food pellets when a criterion was reached. The monkeys in this task tended to choose the 'high' icon after correct responses and the 'low' icon after incorrect responses. Moreover, they did so within the first few trials of transferring to this task (the monkeys had previously been trained to respond metacognitively in other, perceptual, tasks; see Son and Kornell 2005). It seems, then, that they had learned a broad metacognitive skill that could generalize to new circumstances. Crucially, the monkeys appear to have represented two internal responses: a recognition memory response and a confidence judgement, as measured by their wagers. These data do not imply that the monkeys, one of whom was Lashley, by the way, had self-awareness. They do, however, imply that the animals could monitor their confidence in their own memories—a true metacognitive judgement (for recent reviews of animal metacognition research, see Kornell (2009), Smith (2009), and Terrace and Son (2009)).

The ambiguous case of Panzee the chimp: noetic or autonoetic metacognition?

Panzee, a female chimpanzee, had been taught to use over 100 lexigrams, at the time of the 'experiment' in which one keeper hid 26 food objects and seven non-food objects in a large forest field, an area that Panzee knew from her past, but had not visited in 6 years (Menzel 2005). Panzee was able to recruit the assistance of other caretakers (who knew nothing about the objects being hidden) and 'tell' them where the objects were hidden. Because these new caretakers were not aware of the 'experiment' at all, let alone where the objects were hidden, when objects were found, it was thought to be the result of Panzee's 'own initiative' (Menzel 2005, p. 199). The uninformed caretaker found all 34 objects as a result of Panzee's behaviour! And, furthermore, Panzee had indicated on her lexigram board 84% of the time, which particular item had been hidden in each location, and correctly identified these items at delays, for some items, of over 90 hours from the original hiding event. Evidence in support of metacognition was seen in Panzee's behaviour:

The caretaker noted and responded to Panzee's relative degree of excitement—a seemingly spontaneous metacognition, since it directly reflected the distance to the target. Panzee kept pointing, showed intensified vocalization, shook her arm, and bobbed her head or body as the caretaker got closer to the site (see Menzel 2005, p. 202). In addition, Menzel reported that Panzee seemed to do whatever it took to catch the caretaker's attention and, only once joint attention was established, touched the lexigram corresponding to the type of object hidden, pointed outdoors, sometimes went outdoors (if the caretaker followed), and continued to point manually toward the object and vocalize until the caretaker found the object. As noted by Kohler (1925), the 'time in which chimpanzees live' and whether they are able to freely mentally time travel, as autonoetic consciousness requires, remains an open question, but it seems, from these data that Panzee could, at the very least, freely recall which one of at least 20 types of objects she had been shown at a distance and at a long delay, and that she was highly certain, and highly keyed up, of her own knowledge—a feat that begins to look a lot like human autonoesis.

Autonoetic metacognition: self-referential judgements about internal representations

There are several kinds of metacognitive judgements that seem autonoetic. The criterion is that the judgement be specifically self-referential. The three main categories of research that conform to this definition of autonoetic metacognition are source judgements, remember/know judgements, and agency judgements.

Source judgements

While there is a large literature on source judgements (see Johnson et al. 1993; Mitchell and Johnson 2009), most of that literature is not specifically self-referential. For example, much effort has been invested in determining when and under what circumstances people are able to distinguish one person from another as the source of an utterance, but neither person is the self, or whether the original input was auditory or visual, say, or whether the background colour was red or blue. Young children and older adults (Craik et al. 1990; Henkel et al. 1998) especially have difficulties with source judgements. But none of them qualify as necessarily being autonoetic.

However, certain source judgement are necessarily autonoetic, if the distinction the individual must make involves the self as compared to another, or the self in one form (imagining speaking, say) as compared to in another form (actually speaking). People with schizophrenia have particular difficulty with this kind of judgement (Wang et al. 2010). Furthermore, deficits in self-other source (but note, these are often not distinguished from non-self-referential source judgements in the literature) appear to be related to positive symptoms of schizophrenia such as hallucinations and delusions.

Many of the results in the source monitoring literature focus on the details of memories of past events, and some of these studies—those that are particularly relevant for self-consciousness—investigate the extent and manner of self-involvement in those memories. However, it could be argued that a simpler kind of metacognition—that involving adjectival checklists, or self referential statements—is also a kind of metacognitive judgement that is also autonoetic. When a person is asked to decide whether they are warm, attractive, miserly, or intelligent, presumably these judgements are specifically referred to a representation of the self, and would need to be called autonoetic by our definition of the term. Interestingly, when one is making such judgements, there is a particular area of the medial prefrontal cortex that appears to be selectively activated (Ochsner et al. 2005; Jenkins and Mitchell 2011). That area is also often found to be activated in episodic memory task that Tulving would call autonoetic in nature—a fascinating relation that

deserves further research. It is conceivable that this area is, in some sense that is undoubtedly too simple but nevertheless intriguing, the seat of the self.

Remember–know judgements

Judgements concerning whether the individual remembers that an event happened in his or her personal past, or just knows that something is familiar (Tulving 1985; Gardiner 1988) are metacognitive judgements proper, that, taken at face value, are specifically self-referential and hence autonoetic (Gardiner et al. 1998; Hirshman 1998; Yonelinas 2002). Indeed, they have often been taken as the most quintessential of autonoetic judgements.

There is, however, dispute in the literature about exactly how the individual makes remember-know judgements. If they simply evaluate the amount of information that can be retrieved, and say that they 'remember' when they have retrieved a great deal of information, and that they 'know' when they have retrieved a lesser amount of information, then these judgements are essentially retrospective confidence judgements. As with confidence judgements detailed in the previous section, they would be noetic rather than autonoetic judgements. Some researchers have argued for such an explanation, demonstrating that many of the characteristics of remember/know judgements can be handled within a signal detection framework (Donaldson 1996; Dunn 2004; Wixted and Stretch 2004). However, Yonelinas (2002) and others (e.g. Wolk et al. 2006) have argued that two processes are involved: familiarity monitoring and recollective retrieval. These dual process theorists get closer to the original idea that there is something special and different about 'remember' judgements. But even in this dual process view, the more complex form of memory access (i.e. recollective retrieval) is not necessarily self-referential. Insofar as the judgement that one remembers *is* self-referential, then, the remember–know paradigm would appear to be an autonoetic form of metacognition, but neither model of the task emphasizes this characteristic.

Agency judgements

People are able to make fairly reliable judgements of their own agency—they can assess the extent to which they were or were not the causal agent in producing an action outcome (Metcalfe et al. 2010; Miele et al. 2011), a clearly self-referential metacognition. However, they cannot do so infallibly. Wegner and Wheatley (1999; Wegner 2003; Wegner et al. 2004) have provided several fascinating experimental examples of errors in these judgements. In one study, participants, wearing headphones, with their hands at their sides, looked at a mirror image of themselves covered by a smock with the hands of a confederate protruding where their own hands would normally be seen. The participants, of course, knew that the hands that they were seeing in the mirror were not their own hands. But if a word for an object was primed (via the headphones) at just the right moment before the hands that looked like their own hands moved, people had a spooky feeling that they had reached for the object. Their judgement of agency, hence, was malleable and subject to illusion.

But while agency judgements can be distorted (as can lower-level metacognitions), they are normally accurate. For example, Metcalfe and Greene (2007) showed that college students usually correctly know when they have moved a mouse to catch a target, and when noise-like interference, which distorted their own planned movements, intervened. Knoblich et al. (2004) showed that while typical adults can detect a distortion in their motor movements, patients with schizophrenia have great difficulty in doing so.

What about non-human animals? The data, so far, are scant but promising on this issue (Couchman 2012). But, insofar as one component of metacognitive judgements of agency involves action monitoring non-human primates may—given their dexterity and physical

competence—be excellent at it. Originally the comparator action monitoring models (Wolpert et al. 1995), that form the core of most theoretical views of how people make judgements of agency, were devised as a way of understanding how it is possible for people to make nuanced and complex fast actions. The central idea is that the person has a plan of where and how to move. This plan runs off mentally in real time synchronously with their actual movement, and the feedback from the movement is collated with the expectations from the plan. If the two correspond perfectly, the action proceeds smoothly. If there is a mismatch, then an alteration is needed to correct the movement. This match/mismatch mechanism, devised for motor control, was co-opted by the metacognitive system, to allow people to make judgements of agency: if there is no discrepancy, then the person was in control. If a discrepancy occurred, though, then some outside source was distorting the correspondence between intent and action, and the person was not in full control. Presumably to accomplish acrobatic feats so common in the wild, our primate ancestors would need to have a finely tuned action monitoring system. Whether, like humans, they co-opted it to allow them to have metacognition of agency and perhaps even a concept of the self, we do not know.

Conclusion

Is it conceivable that a non-human animal or a computer could exhibit autonoetic metacognition? So far, to our knowledge, no computer has ever done any truly self-referential task. But typically, computers are not programmed to remember their past or project into their future. Nor are they programmed to take particular account of things they themselves did. But there seems to be no 'in principle' reason why this could not be programmed into them. It is imaginable that a computer-robot could be programmed to encode the visual scenes that occurred from their perspective while they moved around in the environment and use those 'personal' records in later encounters, tagging particular knowledge as specific to them. Watson, too, could be programmed to tag his own answers and those of the other participants such that he could later 'remember' the source of the answers. But if that were done would it mean that Watson would have autonoetic metacognition?

One argument against this is that, although such noting and tagging would allow him to give answers that mimicked those of a person who had a self, the records of the computer would comprise a pseudo self. Humphrey (2006) has made a fascinating case that the internalized concept of a self developed in animals because it bestowed evolutionary advantages on those who had it. The advantage accrues because the self as an embodied and encapsulated concept results in an individual who both has a mind, and has a concept of its own physical body and, thereby, strives to preserve and foster it. If one compared an animal with a self to one without, the former would be more motivated to protect its physical body. And, of course, protecting one's body is evolutionarily advantageous. If the 'real' self is necessarily linked to some such creature-based evolutionary account, then even if Watson could access the digital records taken from his perspective, or could answer Watson versus other source questions correctly, he would not thereby manifest a 'real' self. The deep and meaningful characteristics of what self-reference means to humans and to their survival would not follow from answering such questions correctly. In short, the answers to the questions directed at determining whether the answerer has autonoetic consciousness could be faked.

How does metacognition relate to self-awareness, then? First of all, we have argued that anoetic and noetic metacognition do not imply self-awareness at all. That being the case, even humans may not always be self-aware when making metacognitive judgements (e.g. Son and Kornell 2005). But autonoetic metacognition (as long as it is not faked) suggests that the individual has

self-awareness, and an internalized, articulate concept of the self. Now, of course, humans may also be self-aware at other times—the argument is only that anoetic and noetic metacognition provide no positive evidence.

At present, we know almost nothing about self-awareness in non-human primates and other animals. The question has not yet been posed. But, if someone were able to convincingly devise a method of asking a monkey whether he was the agent or someone else was, he might be able to answer it correctly. And, it would not be too far fetched to suppose that—in the complex social world in which primates in the wild live, in which keeping track, over time, of exactly who did what to whom might enhance one's chances of survival—a self might be a valuable thing to have.

References

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Couchman, J. J. (2012). Self-agency in rhesus monkeys. *Biological Letters*, 8(1), 39–41.

Craik, F. I., Morris, L. W., Morris, R. G., and Loewen, E. R. (1990). Relations between source amnesia and frontal lobe functioning in older adults. *Psychology & Aging*, 5, 148–51.

Descartes, R. (1637). *Discourse on Method*. Cambridge University Press.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, 24, 523–33.

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review*, 111, 524–42.

Finn, B. and Metcalfe, J. (2010). Scaffolding feedback to maximize long term error correction. *Memory & Cognition*, 38, 951–61.

Foote, A. L. and Crystal, J. D. (2007). Metacognition in the rat. *Current Biology*, 17, 551–5.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, 16, 309–13.

Gardiner, J. M., Richardson-Klavehn, A., and Ramponi, C. (1998). Limitations of the signal-detection model of the remember–know paradigm: A reply to Hirshman. *Consciousness & Cognition*, 7, 285–8.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9), 5359–62.

Hawkins, S. A. and Hastie, R. (1990). Hindsight: Biased judgements of past events after the outcomes are known. *Psychological Bulletin*, 107, 311–27.

Henkel, L. A., Johnson, M. K., and De Leonardis, D. M. (1998). Aging and source monitoring: Cognitive processes and neuropsychological correlates. *Journal of Experimental Psychology: General*, 127, 251–68.

Hirshman, E. (1998). On the utility of the signal detection model of the remember–know paradigm. *Consciousness & Cognition*, 7, 103–7.

Hoffrage, U. and Pohl, R. (2003). Research on hindsight bias: A rich past, a productive present, and a challenging future. *Memory*, 11, 329– 35.

Humphrey, N. (2006). *Seeing red: A study in consciousness*. Boston, MA: Harvard University Press.

Husserl E. (1929). *Cartesian Meditations and the Paris Lectures*. The Hague: Martinus Nijhoff, (1973).

James, W. (1890). *The Principles of Psychology, Volume 1*. New York: Henry Holt.

Jenkins, A. C. and Mitchell, J. P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience*, 6(3), 211–18.

Johnson, M. K., Hashtroudi, S., and Lindsay, D. S. (1993). Source monitoring. *Psychological Bulleting*, 114, 3–28.

Kahneman, D., Knetsch, J., and Thaler, R. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, 98, 1325–48.

Knoblich, G., Stottmeister, F., and Kircher, T. T. J. (2004). Self-monitoring in patients with schizophrenia. *Psychological Medicine*, 34, 1561–9.

Kohler, W. (1925). *The Mentality of Apes*. New York: Routledge and Kegan Paul.

Koriat, A., Bjork, R. A., Sheffer, L., and Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, 133, 643–56.

Kornell, N. (2009). Metacognition in humans and animals. *Current Directions in Psychological Science*, 18, 11–15.

Kornell, N., Son, L. K., and Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18, 64–71.

Menzel, C. R. (1999). Unprompted recall and reporting of hidden objects by a chimpanzee (*Pan troglodytes*) after extended delays. *Journal of Comparative Psychology*, 113, 426–34.

Menzel, E. (2005). Progress in the study of chimpanzee recall and episodic memory. In H. S. Terrace and J. Metcalfe (Eds.) *The missing link in cognition*, pp. 188–224. Oxford: Oxford University Press.

Metcalfe, J. and Greene, M. J. (2007). Metacognition of agency. *Journal of Experimental Psychology: General*, 136, 184–99.

Metcalfe, J. and Kober, H. (2005). Self-reflective consciousness and the projectable self. In H. S. Terrace and J. Metcalfe (Eds.) *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*, pp. 57–83. Oxford: Oxford University Press.

Metcalfe, J., Eich, T. S., and Castel, A. (2010). Metacognition of agency across the lifespan. *Cognition*, 116, 267–82.

Miele, D. M., Wager, T. D., Mitchell, J. P., and Metcalfe, J. (2011 Dissociating neural correlates of action monitoring and metacognition of agency. *Journal of Cognitive Neuroscience*, 23(11), 3620–36.

Mitchell, K. J. and Johnson, M. K. (2009). Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory? *Psychological Bulletin*, 135, 638–77.

Ochsner, K. N., Beer, J. S., Robertson, E. R., *et al.* (2005). The neural correlates of direct and reflected self-knowledge. *Neuroimage*, 28, 797–814.

Nelson, T. O. and Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.) *The psychology of learning and motivation* (Vol. 26), pp. 125–41. New York: Academic Press.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology*, 19, 90–138.

Reder, L. M. and Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.) *Implicit Memory and Metacognition*, pp. 45–77. Mahwah, NJ: L. Erlbaum.

Russell, B. (1997). *The Problems of Philosophy*. New York: Oxford University Press.

Rosenbaum, R. S., Köhler, S., Schacter, D. L., et al. (2005). The case of K.C.: Contributions of a memory-impaired person to memory theory. *Neuropsychologia*, 43, 989–1021.

Rosenthal, D. (2000) Consciousness, content, and metacognitive judgements. *Consciousness Cognition*, 9, 203–14.

Schwartz, B. L. and Metcalfe, J. (2011). Tip-of-the-tongue (TOT) states: Retrieval, behavior, and experience. *Memory & Cognition*, 39(5), 737–49.

Shields, W. E., Smith, J. D., Guttmannova, K., and Washburn, D. A. (2005). Confidence judgments by humans and rhesus monkeys. *Journal of General Psychology*, 132, 165–86.

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Smith J. D. (2009). The study of animal metacognition. *Trends in Cognitive Science*, 13, 389–96.

Smith J. D., Schull, J., Strote, J., McGee, K., Egnor, R., and Erb, L. (1995). The uncertain response in the bottlenosed dolphin (*Tursiops truncatus*). *Journal of Experimental Psychology: General*, 124, 391–408.

Smith, J. D., Shields, W. E., Allendoerfer, K. R., and Washburn, W. A. (1998). Memory monitoring by animals and humans. *Journal of Experimental Psychology: General*, 127, 227–50.

Smith, J. D., Beran, M. J., Couchman, J. J., Coutinho, M. V. C., and Boomer, J. B. (2009). Animal metacognition: Problems and prospects. *Comparative Cognition and Behavior Reviews*, 4, 40–53.

Son, L. K. and Kornell, N. (2005). Meta-confidence judgments in rhesus macaques: Explicit versus implicit mechanisms. In H. S. Terrace and J. Metcalfe (Eds.) *The Missing Link in Cognition: Origins of Self-Knowing Consciousness*, pp. 296–320. New York: Oxford University Press.

Son, L. K. and Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage model. *Memory & Cognition*, 33, 1116–29.

Staddon, J. E. R., Jozefowiez, J., and Cerutti, D. T. (2007). Metacognition: A problem not a process. *PsyCrit*.

Terrace, H. S., and Son, L. K. (2009). Comparative metacognition. *Current Opinion in Neurobiology*, 19, 67–74.

Thaler, R. (1980). Towards a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, 1, 39–60.

Tulving, E. (1984). Elements of episodic memory. *Behavioral and Brain Sciences*, 7, 223–68.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26, 1–12.

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? In H. S. Terrace and J. Metcalfe (Eds.) *The Missing Link in Cognition: Origins of Self-Knowing Consciousness*, pp. 4–56. New York: Oxford University Press.

Wang, L., Metzak, P. D., and Woodward, T. S. (2010). Aberrant connectivity during self-other source monitoring in schizophrenia. *Schizophrenia Research*, 125, 136–42.

Washburn, D. A., Gulledge, J. P., Beran, M. J., and Smith, J. D. (2010). With his memory erased, a monkey knows he is uncertain. *Biology Letters*, 6, 160–2.

Wegner, D. M. (2003). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wegner, D. M. and Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54, 480–92.

Wegner, D. M., Sparrow, B., and Winerman, L. (2004). Vicarious agency: Experiencing control over the movements of others. *Journal of Personality and Social Psychology*, 86, 838–48.

Wheeler, M. A., Stuss, D. T., and Tulving, E. (1997). Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin*, 121, 331–54.

Wixted, J. T. and Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomics Bulletin & Review*, 11, 616–41.

Wolk, D. A., Schacter, D. L., Lygizos, M., *et al.* (2006). ERP correlates of recognition memory: Effects of retention interval and false alarms. *Brain Research*, 1096, 148–62.

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995), An internal model for sensorimotor integration. *Science*, 269, 1880–2.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory & Language*, 46, 441–517.

Zalla T., Stopin, A., Ahade, S., Sav, A. M., and Leboyer, M. (2008). Faux-pas detection and intentional action in Asperger syndrome. A replication on a French sample. *Journal of Autism and Developmental Disorders*, 39, 373–82.

Chapter 19

Seeds of self-knowledge: noetic feelings and metacognition

Jérôme Dokic

Feeling is to knowledge what a cry is to a word.
Erwin Straus

Introduction

As authors from various traditions and disciplines—including phenomenology, cognitive and social psychology—have observed, our most spontaneous judgements can reflect what we ordinarily call 'our feelings'. Sometimes we judge that something is the case just because (*ceteris paribus*) we *feel* that this is so. Feeling-based judgements seem to provide us with information that it would have been difficult, perhaps impossible, to acquire through other epistemic means, such as perception, memory, and inference. As a consequence, they can act as first premises in both theoretical and practical reasoning. In many everyday circumstances, we are ready to judge, reason, and act on the basis of our feelings without further ado.

If ordinary language descriptions of our feelings are to be trusted, the latter can be about external states of affairs ('I feel that it's going to rain'), as well as about our own bodily states and dispositions ('I feel tired', 'I feel elated'). In this chapter, though, I am interested in another species of feelings, namely those that concern our own mental and epistemic life. I shall call the relevant feelings 'noetic feelings'; they have also been called 'epistemic' or 'metacognitive' feelings.[1] Here is a partial and non-exhaustive list of noetic feelings as they have been discussed in the literature:

- *Feelings of knowing/not knowing* (Koriat 1995, 2000).
- *Tip-of-the-tongue experiences* (Brown 2000; Schwarz 2002).
- *Feelings of certainty/uncertainty* (Smith et al. 2003).
- *Feelings of confidence* (Winman and Juslin 2005).
- *Feelings of ease of learning* (Koriat 1997).
- *Feelings of competence* (Bjork and Bjork 1992).
- *Feelings of familiarity* (Whittlesea et al. 2001a, 2001b).
- *Feelings of 'déjà vu'* (Brown 2003).

[1] See Koriat (2006, p. 54), who writes that there is an 'assumption underlying much of the work in metacognition […], that metacognitive feelings play a causal role in affecting judgments and behavior'.

- *Feelings of rationality/irrationality* (James 1879).
- *Feelings of rightness* (Thomson 2008).

These feelings are noetic in the sense that they intuitively concern epistemic states, events, or skills, although the sense in which this is so needs careful delineation. Admittedly, the boundary between noetic feelings and other kinds of feelings is not very sharp. Some feelings seem to lie at the borderline between noetic feelings and feelings about the external world. For instance, it is not clear whether the feeling of presence (Matthen 2005) is just the feeling that a state of affairs is actual (rather than merely possible), or the feeling that one is genuinely *related* to the actual world. Similarly, the feeling that something in the visual field has changed (Rensink 2004; Loussouarn 2010) might really be the feeling that one has *detected* a change, even though one is not able to identify it. In advance of a substantial theory of feelings, it is hard to classify these feelings as genuinely noetic or not. In any case, I shall focus here on feelings which are clearly noetic, such as the feeling of knowing and the feeling of (subjective) uncertainty.

This chapter is structured as follows. In the first section, I discuss a concrete example illustrating the fact that noetic feelings are 'seeds' of self-knowledge, i.e. can provide knowledge or justified beliefs about one's own mental and epistemic life. Then, in the next three sections, I formulate three theoretical models of the psychological nature and epistemic value of noetic feelings. On the Simple Model, noetic feelings are manifestations of metarepresentational states of knowledge that are already in place. On the Direct Access Model, they are (possibly partly opaque) experiences about one's own first-order states of knowledge. Finally, on the Water Diviner Model, they are first and foremost bodily experiences, whose objects (bodily states) are only contingently associated with first-order epistemic states. Still, they can acquire a derived content representing or concerning such states. The latter model will turn out to be superior to the other ones. First, it helps to disambiguate the sense in which noetic feelings can be described as 'metacognitive' ('Metacognition versus metarepresentation' section). Second, it can easily be extended to deal with the motivational dimension that many noetic feelings seem to have ('Noetic feelings and motivation' section). In the following section ('Two kinds of metacognition, and a case study'), I build on the account sketched in the previous sections and illustrate the distinction between two kinds of metacognition (which I call 'procedural' and 'deliberate') with respect to feelings of uncertainty experienced in the context of certain perceptual categorization tasks. Eventually, in the section entitled 'The Competence View', I put forward a tentative hypothesis about the derived intentional contents of noetic feelings, according to which they can concern our own mental and epistemic life without being strictly speaking metarepresentational, i.e. without being constitutively linked to the possession of metarepresentational or mindreading abilities.

Feelings of knowing and self-knowledge

Consider the following pair of questions:

Q1 Is Lima the capital of Peru?

Q2 Do you believe that Lima is the capital of Peru?

On the face of it, these are very different yes–no or polar questions, despite the fact that they have overlapping contents. Q1 is a question about the geographical world, whereas Q2 is a question about the addressee, more precisely about whether she is in a specific mental state, namely the state of believing that Lima is the capital of Peru. Yet the answer to Q2 can be directly based on an answer to Q1. The addressee can answer 'yes' to Q2 if she is ready to answer 'yes' to Q1. Indeed, if she fully understands both questions, she normally *cannot* answer 'yes' to Q2 without thereby being in a position to answer 'yes' to Q1.

Gareth Evans has drawn the connection between these two types of questions in the following general terms:

> I get myself in position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*. (Evans 1982, p. 225.)

In a later essay, Gordon (1995) calls the procedure that Evans is describing here an 'ascent routine':

> Because this procedure answers a metacognitive question by answering a question at the next lower semantic level, I will call it an *ascent routine*. (Gordon 1995, p. 60.)

Both Evans and Gordon take ascent routines to be *alternatives* to traditional introspective methods. In answering Q2, the addressee does not have to search her mind for a specific belief, much less a state of knowledge. Rather, she directs her attention to the outer world as she conceives it. No introspective ability needs to be invoked in order to determine whether she believes that Lima is the capital of Peru.

Now consider another pair of questions:

Q3 What is the capital of Peru?

Q4 Do you know what the capital of Peru is?

Q3 and Q4 are very different non-polar questions, despite the fact that they have overlapping contents. The former is about the geographical world, whereas Q4 is a question about the addressee. Yet the addressee *can* answer Q4 (by saying 'yes') without being in a position to answer Q3 (by saying 'yes, Lima'). In fact, she can answer Q4 without having any city in mind.

There are two ways she can do this. One way is to use independent information to the effect that she is competent in answering a first-order question such as Q3. For instance, she knows that she was a good geography student at school, and that she learnt all the capitals in the world by heart. In such a case, her metacognitive judgement to the effect that she can answer Q3, on which she can ground a 'yes' answer to Q4, is *theory-based*. It inferentially derives from independent beliefs based on memory. Alternatively, the addressee may just *feel* that she knows what the capital of Peru is. She feels competent in answering Q3, in advance of actually providing any answer, either privately or publicly. In this case, her metacognitive judgement is *experience-based*. It seems to be directly based on her affective experience (a 'gut feeling') independently of background beliefs.[2]

What is the nature of the feeling of knowing which enables one to answer a question such as Q4 in advance of giving any answer to Q3? In particular, since ascent routines are clearly not available in this case, is such a feeling a form of introspection of one's own epistemic states? In what follows, I shall present three models of feelings of knowing that try to provide answers to these questions.

The Simple Model

On the Simple Model, noetic feelings are in fact metarepresentational beliefs, more precisely beliefs that are explicitly about one's epistemic states (Dienes and Perner 1999). For instance, the feeling that the subject knows the name of the capital of Peru is just the actualization of a piece of higher-order knowledge acquired long ago, namely the knowledge that she knows the name of the capital of Peru. Of course, if she is wrong and in fact she does not know that the capital of Peru is

[2] The distinction between theory-based (or information-based) and experience-based metacognitive judgements comes from Koriat (2006).

called 'Lima', her feeling expresses mere apparent knowledge, but it is still the actualization of a higher-order mental state, more precisely a false belief about her first-order state of knowledge.

The Simple Model can thus provide a straightforward explanation of why we can have feelings of knowing while being actually unable to retrieve the relevant name, as it happens in so-called 'tip-of-the-tongue' experiences. Surely, the higher-order state of knowledge, or apparent knowledge, that we know the name of the capital of Peru can be made explicit while the corresponding first-order state of knowledge, or apparent knowledge, that the capital of Peru is called 'Lima' remains implicit because of some performance problem. These can be distinct states, and either one can be activated independently of the other. In the case of geographical ignorance, a higher-order state of apparent knowledge that we know the name of the capital of Peru can even exist in the absence of any first-order state of knowledge to the effect that Lima is the capital of Peru.

I call this model 'simple' because it does not posit new kinds of mental states, since noetic feelings are assimilated to ordinary beliefs, in the form of higher-order memory states. On this model, noetic feelings can justify other beliefs because they are themselves beliefs. Besides, we often lose the original justification of our memory beliefs, a fact that might be invoked in order to explain why we are not fully aware of the underlying reasons for what our feelings tell us. Despite its relative simplicity, though, the Simple Model faces several difficulties.

The first difficulty will become clearer as we proceed. It concerns the fact that on the Simple Model, noetic feelings necessarily have metarepresentational contents. They are explicitly about first-order states of knowledge. It follows that the subject must possess relevant epistemic concepts, such as the concept of knowledge or memory, in order to *have* noetic feelings. In order words, noetic feelings are available only to creatures possessing a theory of mind. However, as we shall see (see especially the last two sections), there are reasons to think that creatures lacking metarepresentational resources can still have noetic feelings, such as feelings of knowing and feelings of uncertainty, and exploit them in theoretical and practical reasoning.

Another difficulty is that even if the subject has metarepresentational abilities, noetic feelings seem to be sources of *original* knowledge or justified beliefs, at least in some cases. After all, perhaps the subject never acquired the higher-order knowledge that she knows the name of the capital of Peru, or she might have forgotten about it a long time ago. Still, she can have the feeling that she has such knowledge just because she is being asked a question such as Q3 ('What is the capital of Peru?'). In this case, it seems that her feeling of knowing enables her, in concert with the fact that she possesses the relevant mental concepts, to acquire a new piece of higher-order knowledge. In contrast, if feelings of knowing are already conceived as higher-order beliefs, it is not clear that they can be justified or warranted.

Finally, the Simple Model forces its proponents to adopt a curious interpretation of well-replicated experimental results. It appears that feelings of knowing can be easily manipulated in certain experimental conditions (see, e.g. Reder 1987; Bjork 1999). For instance, by *priming* some of the question terms, psychologists can raise the feeling of familiarity toward a question such as Q3, and produce a fairly convincing feeling that the subject knows the answer to the question, even if she does not. On the Simple Model, these experimental manipulations must be interpreted as creating *false* higher-order memories in the subject, which is quite implausible, at least on the face of it.

The Direct Access Model

On the Direct Access Model, noetic feelings are cases of *introspection*. They provide us with internal awareness of our own first-order memories as carrying information relevant to answering certain questions. So when the subject feels that she knows the name of the capital of Peru, she has

in fact access to one of her first-order states of knowledge, namely the memory that the capital of Peru is called 'Lima'. In the case in point, the subject is not conscious of her memory as having the content 'The capital of Peru is called "Lima"'. Rather, she is conscious of her memory only as having a content of the form 'The capital of Peru is called ____'. In other words, she has introspective access to her memory as such while having access only to a *proper part* of its content.[3] Of course, if the subject does not really know that the capital of Peru is called 'Lima', her feeling of knowing is somehow non-veridical. Still, in this case, she has the *apparent* introspective experience of having the relevant information stored in her mind.

Unlike the Simple Model, the Direct Access Model can explain why noetic feelings are, at least sometimes, a source of original knowledge or justified beliefs about our mental states and dispositions. The subject's feeling of knowing can reveal a piece of information about herself that she may never have explicitly acquired before, namely that she possesses information relevant to answering a question such as Q3.[4] Noetic feelings belong to a class of *experiential* states, so that beliefs based on them can act as bona fide premises in theoretical and practical reasoning. In other words, these beliefs are justified by a belief-independent affective experience, just as perception or memory beliefs are justified by belief-independent perceptual or memory experiences.

It is helpful to compare the Direct Access Model with David Rosenthal's analysis of the tip-of-the-tongue experience:

> When I have Mark Twain's real name on the tip of my tongue, I must be conscious *of* the particular state that carries that information. But I am not conscious of that state in respect of the specific information the state carries; rather, I am conscious of the state only *as* a state that carries that information. (Rosenthal 2000, p. 204.)

Rosenthal draws a distinction between being conscious of a given informational state (for instance, the memory that Mark Twain's real name is 'Samuel Clemens') in respect of the specific information the state carries and being conscious of it only in respect of what questions the information would answer. However, Rosenthal does not defend the Direct Access Model, because he makes clear that being conscious of a given informational state only in respect of what questions the information would answer does *not* entail that this state is itself a conscious state. In contrast, at least to the extent that the objects of conscious introspection must themselves be conscious states, the Direct Access Model entails that feelings of knowing are ways of bringing to consciousness relevant informational states, even though their contents are at least partly occluded to the subject.

Of course the Direct Access Model is hostage to a substantial theory of introspective knowledge, and in particular to the issue of whether the latter should be conceived on the model of observational knowledge. Independently of this issue, though, it is important to notice that the Direct Access Model, at least as applied to feelings of knowing, is incompatible with two general views about introspective knowledge. The first view is that introspection makes the subject aware of her own intentional mental states only by revealing their contents (see, e.g. Tye 2009). In other words,

---

[3] The Direct Access Model is not committed to the claim that all types of noetic feelings involve opacity in this sense. Certainly feelings of knowing are not unique in this respect. For instance, on this model, the feeling of familiarity relative to a particular perceived person would be the introspective experience of memories involving this person, but whose contents are at least partly opaque to the subject. In other words, the subject knows that she has memories about the person while being temporarily unable to access the full contents of these memories.

[4] So the subject knows that she is competent in answering certain questions *in virtue* of the fact that she is aware of one of her memories as carrying information of a certain kind.

introspection is *fully transparent* with respect to the contents of the introspected states (whenever they have contents). The Direct Access Model denies that introspection is always transparent in this sense, since feelings of knowing are precisely introspective states about particular first-order memories, while their contents are only partially revealed to the subject.

Another, less radical view of introspection or self-knowledge that is incompatible with the Direct Access Model is the 'hierarchy of explicitness' view (as we may call it) according to which the awareness of the contents of one's own mental states is a *precondition* of the awareness of the fact that one is in them (Dienes and Perner 1999, 2002). Unlike the first view, this view acknowledges that introspection can reveal the *mode* of the introspected state, but only if the latter's content has already been fully revealed to consciousness. In contrast, the Direct Access Model allows for a mode to be revealed by introspection (in the case in point, the fact that the introspected state is a *memory*), while only revealing part of the introspected state's content.

At this stage, the Direct Access Model might seem to be a more serious competitor than the Simple Model. Still, the empirical evidence is not quite favourable to it. Psychological experiments suggest that what determines feelings of knowing need *not* be familiarity with the answer. Rather, at least some feelings of knowing are determined by familiarity with question terms (Reder and Ritter 1992) and/or accessibility of partial information regardless of its adequacy (Koriat and Levy-Sadot 2001). In other words, the implicit mechanisms underlying the feeling of knowing need not monitor the memory trace itself (*pace* Hart 1965). In fact, they can be causally disconnected from the subject's first-order state of knowledge. Insofar as the notion of sensitivity is a causal-informational one, they are not sensitive (they do not have direct access) to the presence in long-term memory of the name to be retrieved.[5]

It follows that a natural causal explanation of introspective awareness is not available to proponents of the Direct Access Model. According to this explanation, a necessary condition of being introspectively aware of a mental state M is that M be the *cause* of one's introspective awareness. However, empirical evidence suggests to the contrary that feelings of knowing are not caused by first-order memory states (or corresponding memory traces in the brain), but rather by cues (processing fluency, availability of partial information) that are only contingently associated with these states, which might not even exist. Now whether this is incompatible with the claim that feelings of knowing involve a form of direct introspective access to one's own mental states at the personal level remains to be determined.

The Water Diviner Model

The Water Diviner Model is named after a character introduced by Wittgenstein in *The Blue Book*, who claims to *feel* (the German verb is 'fühlen') in his hand that there is water three feet underground. On this model, noetic feelings are first and foremost bodily experiences, i.e. experiences about bodily states. They are diffuse affective states registering internal physiological conditions and events. Unlike bodily sensations, though, noetic feelings need not have precise locations in external bodily parts. At a phenomenological level, they often have an 'indistinct, spreading, blurred quality' and they 'seem to actively resist attempts to focus attention directly on them' (Mangan 2001). In William James's terms, they belong to the 'fringe of consciousness' (James 1890).

Now, just as the water diviner's sensations reliably co-vary with physical conditions, namely the presence of water underneath, noetic feelings reliably co-vary with *mental* conditions.

[5] Of course, other types of noetic feelings may be such that their underlying metacognitive mechanisms are causally sensitive to the relevant target in memory. Metcalfe (2000) argues that this is the case with 'feelings of imminence', such as those involved in tip-of-the-tongue experiences.

For instance, the feeling of knowing co-varies with the fact that the subject has knowledge about the relevant subject-matter. As a result, at least some particular feelings of knowing indicate or carry information about the presence of first-order states of knowledge. In other words, feelings of knowing 'track' such states, in the sense that normally, the former occur only in the context of the latter ('I would not have the feeling of knowing this person's name if I did not know it'). The cues underlying noetic feelings are contingently but stably associated with epistemic states. This association holds in a normal (ecological) context, but it can be severed by psychologists, who can easily produce 'illusory' feelings of knowing (Bjork 1999).

The informational properties of many token feelings can be exploited by a sophisticated cognitive system to recruit types of feelings as representations of mental states. In other words, there is room for an account of noetic feelings that is analogous to familiar teleological-functionalist accounts of emotions. For instance, Prinz (2004, 2007) argues that emotions are perceptions of bodily states that are recruited to represent core relational themes or concerns, such as danger or loss. In his terminology, the 'nominal' contents of emotions are bodily changes, but the 'real' contents of emotions are core relational themes. Similarly, one may argue that the nominal contents of noetic feelings are bodily changes, but the real contents of feelings are mental states.

However, the analogy between noetic feelings and emotions breaks down at a crucial point. The association between basic emotions and their real contents is robust, and possibly innate. It is difficult to imagine fear that does not have the function of detecting danger. In contrast, many noetic feelings seem to be recruited by the organism through some form of learning.[6] As an illustration, consider Harris et al.'s (1981) findings. Both 8- and 11-year-old children read anomalous sentences in a story more slowly. However, only the older group is able to pick out the anomalous lines as not fitting the story. According to the authors' interpretation, both groups of children generate 'internal signals' of comprehension failures, but only the older children have learned to *exploit* such signals to locate the *source* of their feelings of difficulty.

These results suggest that the *same* type of noetic feelings (in the case in point, feelings of difficulty or easiness), individuated in bodily terms, can have additional, *acquired* contents that can be exploited in judgements.[7] In the case of organisms possessing metarepresentational abilities, these acquired contents can be explicitly about their own mental states. For instance, feelings of knowing can be recruited as feelings *that* one knows something, by deploying the mental concept of knowledge. It remains an open issue whether noetic feelings can have acquired contents that somehow hinge on the presence of mental conditions but *without* representing them as such. (See the following sections for further discussion of this point.)

According to the Water Diviner Model, feelings have intentional contents beyond the body, but only in a derived way, through some kind of learning or association process. Such a process generates new heuristics, i.e. cognitive shortcuts that enable us to move spontaneously from our feelings to judgements concerning the task at hand. One form that such heuristics can take is that of answering for oneself the question 'How do I feel about it?' in order to simplify a task that is

[6] See Proust (2008). I do not deny that non-basic emotions, such as respect or pride, need to be trained. It is an interesting question whether there is anything like the distinction between basic and non-basic emotions in the case of epistemic feelings, but here I shall leave this question open.

[7] Another interpretation of the results is that the younger children lack the feelings that older children have and exploit. But certainly, the former *behave* as if they felt the difficulty of certain passages, which they spontaneously read more slowly. What this suggests is that feelings of difficulty already involve some low-level metacognitive control, which falls short of the ability to exploit these feelings at the level of explicit reasoning.

particularly complex and demanding (Schwartz and Clore 1996).[8] In the specific case of noetic feelings, the relevant heuristics enable the subject to form non-inferential judgements about her own mental states, such as the judgement that she knows how to answer the question she is being asked.[9]

In some cases, the association between noetic feelings and their 'real' contents can be easily broken. According to Reber et al. (2004), the judgement that a picture is likeable can be based, *ceteris paribus*, on positive affect elicited by processing fluency. Now in the experiments of Winkielman and Fazendeiro (in preparation), some participants were informed that factors having nothing to do with the pictures, such as background music, might influence their feelings toward the pictures. These participants actually cease to experience the pictures as likeable (or likeable to the same extent), undermining the connection between positive affect and positive aesthetic judgement.

In other cases, the heuristics underlying the formation of feeling-based judgements are more robust, and might exhibit modularity effects. For instance, I can get the feeling that I know the person in front of me despite of the fact that I independently know (e.g., from reliable testimony) that my feeling is misleading; I do not know this person at all. Still, the cognitive impression that I know her might persist, at least for a while. However, although feelings can be synchronically modular in this sense, depending on the robustness of the relevant heuristics, they are certainly not diachronically modular. It is possible in principle that noetic feelings lose their contents and acquire different ones, as new heuristics are implicitly learned.

Metacognition versus metarepresentation

I have presented three models of the psychological nature and epistemic value of noetic feelings, focusing on the case of feelings of knowing. Even though it is possible that the Simple Model and the Direct Access Model have some validity with respect to particular cases of noetic feelings, the Water Diviner Model seems to have the widest domain of application. It does not face important objections like its competitors, and it is empirically plausible. In general, the intentionality of noetic feelings beyond the body is not intrinsic but derived. Feelings are intrinsically about the body, but some of them—the noetic ones—can be exploited by the subject as more or less reliable symptoms of the instantiation of mental states or conditions.

The Water Diviner Model acknowledges a distinction between the cognitive processes underlying and grounding noetic feelings and the further, independent cognitive processes that enable the subject to exploit noetic feelings in explicit judgement and reasoning. What I wish to show now is that this distinction helps us to disambiguate the common claim that noetic feelings are 'metacognitive'.

[8] Note that the use of these heuristics involves the self-ascription of feelings as such. This is not the general case. We often move directly from our feelings to metacognitive judgements without going through a representation of feelings as such. Moreover, the Water Diviner Model is compatible with the claim that the process of associating bodily states with specific mental states is coeval with the development of new perceptual-recognitional abilities with respect to the former. In other words, bodily experience itself may be enhanced by the association process.

[9] The notion of non-inferentiality at stake here concerns the personal level. Feeling-based judgements are cognitively spontaneous in something like Bonjour's sense, i.e. they are involuntary, 'coercive,' and not the result of any *introspectible* train of reasoning (Bonjour 1985, p. 117). Of course this is compatible with their being based on subpersonal inferences or computations.

Psychologists usually define metacognition as 'cognition about one's own cognition', or as 'thinking about thinking'.[10] Philosophers, on the other hand, tend to equate metacognition with metarepresentation, i.e. the ability to form representations about other representations, which is usually associated with possessing a mindreading ability or 'theory of mind'.[11] Correspondingly, contents are metarepresentational when they are explicitly about representations as such. For instance, the content of the belief that Pierre believes that it is going to rain is metarepresentational, because of the presence in it of the mental state of *believing* that it is going to rain.

In fact, noetic feelings can be said to be metacognitive in two quite different senses, depending on whether we are talking about their consciously experienced *intentional contents* or their implicit *causal antecedents*.

Firstly, noetic feelings can be said to be metacognitive insofar as their intentional contents yield information (or misinformation) concerning one's own epistemic states, processes, and abilities. The question is whether these contents are also metarepresentational, which would entail that their apprehension required the possession of mindreading abilities. Here we face two alternatives. If we answer 'yes', no creature can exploit noetic feelings in reasoning without deploying some mental concept or proto-concept. For instance, the content of the feeling of knowing a person's name can only be as sophisticated as *that I know this person's name*, which is the representation of a knowledge state as such. In contrast, if we answer 'no', we allow for the possibility that noetic feelings can rationally guide decision-making and the fixation of beliefs in creatures lacking metarepresentational abilities. Of course, the challenge faced by the second alternative is to show that noetic feelings can be self-directed while having first-order contents. As we shall see in a later section ('Two kinds of metacognition, and a case study'), this challenge is highly relevant to the issue of the correct interpretation of important results in the field of animal cognition.

Secondly, the causal antecedents of noetic feelings can be said to be metacognitive insofar as they involve implicit *monitoring* mechanisms that are sensitive to non-intentional properties of first-order cognitive processes. For instance, the feeling of knowing can be based on an implicit evaluation of the *fluency* of the process constituting our spontaneous attempt to remember something. The feeling of familiarity seems to be based on the implicit detection of a discrepancy between expected and actual fluency of processing (Whittlesea et al. 2001a, 2001b). Obviously, the operations of these mechanisms do not require metarepresentational abilities. To begin with, they are sensitive to properties of internal states and processes independently of whatever *contents* they are carrying. If they involve representations of other representations, they do not involve metarepresentations, i.e. representations of representations *as of* representations.[12]

There may be another, more controversial consideration that leads to scepticism about the possibility that implicit metacognitive mechanisms manipulate metarepresentations. One might argue that metarepresentations are necessarily either actually or potentially conscious. There is a constitutive link between the ability to form metarepresentations and the ability to enjoy conscious states. Metarepresentations involve some conception of mental representation, whose complexity makes them available only to conscious creatures and not to sub-personal mechanisms. In contrast, implicit metacognitive mechanisms involve only representations, which

[10] See, for instance, Nelson (1992) and Metcalfe and Shimamura (1994).

[11] A notable exception is Proust (2006, 2007, 2008), who has forcefully and convincingly argued that metacognitive abilities are distinct and independent from metarepresentational abilities.

[12] As Koriat puts it, judgements based on feelings of knowing 'rely on *contentless* mnemonic cues that pertain to the quality of processing, in particular, the fluency with which information is encoded and retrieved' (Koriat 2006, pp. 19–20; my italics).

cannot be or become conscious. As a consequence, they cannot be metarepresentations. They are first-order representations happening to be about internal rather than external states. In a nutshell, they are first-order but self-directed, as opposed to world-directed.

The two senses in which noetic feelings involve metacognitive abilities are largely independent from each other. Even if one acknowledges that the causal antecedents of noetic feelings involve mechanisms that are implicitly sensitive to the quality of first-order processes, the question of whether the intentional contents of noetic feelings can be metacognitive without being metarepresentational remains entirely open. (We shall come back to this question in the section entitled 'The Competence View'.)

Noetic feelings and motivation

Even if the Water Diviner Model is on the right track, it is still incomplete in that it does not deal with an important feature of many types of noetic feelings, namely their *motivational* dimension. Unlike mere intuitions, noetic feelings can intrinsically motivate the subject *to do* something, either at the mental level (e.g., to form a *judgement*) or at the physical level (e.g., to issue a *speech-act* in order to answer a question).[13]

Consider, for instance, tip-of-the-tongue experiences. They are at least partly constituted by a spontaneous *inclination* or *tendency* to search one's memory and retrieve the relevant information (e.g. the proper name that one has on the tip of one's tongue). It is hard to imagine having a tip-of-the-tongue experience in the absence of such inclination. Of course, one may be independently motivated, at a higher level, not to waste too much time on the task at hand, but it may be hard to resist the primitive inclinations provided at a lower level by one's feeling of knowing. Noetic feelings have a quasi-modular motivational dimension, analogous to the quasi-modularity of emotions (de Sousa 1987).

One may hypothesize that the motivational power of noetic feelings *derives* from their causal antecedents, which involve mental events of *trying* to do something. In other words, noetic feelings piggyback on intrinsically motivational states that already fix a (mental and/or physical) goal for the subject.[14]

This hypothesis highlights the Janus-faced character of noetic feelings with respect to behaviour. Noetic feelings both *precede* and *follow* behaviour. On the one hand, noetic feelings precede and causally determine actions, by providing first premises to practical reasoning. For instance, we can exploit a feeling of incompetence relative to a particular test in a practical deliberation over whether we should take the test or not. Let us call 'Type 2' the controlled, deliberate behaviour that can be initiated by noetic feelings. On the other hand, noetic feelings follow or at least accompany inclinations to act that are already in place. For instance, psychological experiments have revealed that the feeling of knowing a person's name can be based on the unconscious feedback from the subject's spontaneous attempt to retrieve the name from memory. We feel that we know the name of the person we are talking to because we are already *trying* to remember it, and perhaps retrieving at least part of the relevant information (such as the fact that the name is

[13] I do not want to claim that all types of noetic feelings have a motivational dimension. For instance, perhaps 'déjà vu' experiences are independent of any inclination to act, physically or mentally.

[14] I assume that the relation between noetic feelings and antecedent behaviour is *causal*, and thus contingent. A stronger assumption is that this relation can be at least partly *constitutive*. On this assumption, at least some noetic feelings *are* in fact bodily facets of tryings.

dissyllabic), even though we cannot consciously access the whole of it.[15] We can call 'Type 1' the spontaneous behaviour that gives rise to noetic feelings.[16]

The fact that noetic feelings follow behaviour is congenial to an analysis of feelings along the lines of the James–Lange theory of emotions (Koriat et al. 2006; Laird 2007). According to this theory, which James contrasted with the commonsensical view that emotions cause behaviour, 'we feel sorry *because* we cry, angry *because* we strike, afraid *because* we tremble' (James 1890, p. 449). When transposed to noetic feelings, the claim is that we have a feeling of knowing *because* we are already trying to retrieve the relevant piece of information (Type 1 behaviour). However, unlike what James assumed in the case of emotions, this claim need not be in conflict with common sense insofar as feelings can also be the starting point of further, Type 2 behaviour.

The motivational character of tryings underlying noetic feelings *constrains* the intentional content of the latter as it is exploited in conscious reasoning. For instance, the feeling of knowing (respectively, the feeling of *not* knowing) is causally based on the subject's trying to remember the name, and partly determines the strategies that should be deployed at the level of practical reasoning, by providing information (or misinformation) to the effect that the relevant name can be found in the subject (respectively, elsewhere, in other more competent persons or in a book). Such pre-established harmony is no mystery as soon as we acknowledge the stepwise character of noetic feelings. It also shows that the derived intentionality of noetic feelings is not as arbitrary as, say, the derived intentionality of language. One cannot interpret noetic feelings in any way we like, on pain of creating behavioural dissonance.

Two kinds of metacognition, and a case study

Let's take stock. What has emerged from the previous two sections is a general distinction between two kinds of metacognition, which I will henceforth call 'procedural' and 'deliberate'. *Procedural metacognition* is constituted by implicit monitoring and control of first-order processes. Procedural metacognition can generate conscious feelings, but the latter remain epiphenomenal in the sense that they do not mediate the interactions between monitoring and control. Feelings are neither causal nor epistemic intermediaries in the processes of procedural metacognition. At the personal level, procedural metacognition appears as a purely practical skill, which manipulates only implicit representations.[17]

Procedural metacognition can be contrasted with *deliberate metacognition*, which enables the rational exploitation of noetic feelings. There is deliberate metacognition when noetic feelings give rise to judgements that can be used in practical and theoretical reasoning. Deliberate metacognition is something that the subject herself does, rather than a mechanism inside her. As we have seen, the question arises whether deliberate metacognition involves metarepresentational

[15] See, for instance, Koriat and Levy-Sadot (2000), Koriat (2006), and Koriat et al. (2006). As Koriat (1995, p. 312) writes: 'It is by attempting to search for the solicited target that one can judge the likelihood that the target resides in memory and is worth continuing to search for'.

[16] The Type 1/Type 2 terminology is of course reminiscent of the System 1/System 2 distinction, which has been used to characterize two systems of reasoning, intuitive and deliberate (see Kahneman and Frederick 2005; Evans and Frankish 2008). However, if Type 2 behaviour is indeed deliberate, I want to leave open here whether Type 1 behaviour necessarily belongs to System 1—perhaps there is also something like monitoring targeted at processes belonging to System 2.

[17] See Reder and Shunn (1996) and Spehn and Reder (2000) for further discussion of the claim that metacognitive monitoring and control need not be mediated by conscious awareness.

abilities or not. So there is in principle a further distinction between two species of deliberate metacognition, one which involves metarepresentations and the other which does not.

A difficult question is whether noetic feelings are *necessarily* based on procedural metacognition. Clearly, many noetic feelings result from the feedback from implicit control processes (Koriat et al. 2006), which are instances of procedural metacognition in the sense just introduced. One might still wonder whether some noetic feelings result from a *dedicated* form of monitoring, i.e. one that enables control only at the conscious, rational level. Although this is not a priori inconsistent, it is empirically doubtful. Given the brain's ability to create cognitive shortcuts, one can surmise that once such a monitoring mechanism is in place, its outputs will soon be exploited directly at the subpersonal level, without the mediation of conscious experience. Thus, it seems to be an empirical fact that deliberate metacognition (whether it takes a metarepresentational form or not) is always based on procedural metacognition, and thus that noetic feelings are essentially motivational in the sense that they reflect behavioural inclinations that are already in place.

In the rest of this section, I would like to apply the distinction between procedural and deliberate metacognition to a case study that comparative psychologists have recently set up. This case study is about another type of noetic feelings, namely feelings of uncertainty as they can arise in some perceptual categorization tasks. Hopefully this will also illustrate the relevance of the distinction for a general theory of noetic feelings.

It has been argued that at least some non-human animals, including dolphins and some species of monkeys, have noetic feelings, such as feelings of uncertainty, which they can use strategically in their reasoning (Smith et al. 2003; Smith 2005, 2009). For instance, in one of David Smith's numerous experiments, a monkey has to touch a visual pattern on the screen when it is judged to be dense, and the symbol 'S' when the pattern is judged to be sparse instead. In another condition, the monkey is also allowed to press a third, so-called 'uncertainty' key, which simply advances it to the next trial. Like human subjects, the monkey can make an adaptive use of the uncertainty key by reducing the number of errors that it would make in a forced-choice condition. Moreover, it uses this key in conditions very similar to those in which human subjects verbally report that they *felt unsure* about the category of the stimulus. Now if monkeys can have feelings of uncertainty, they should have first-order contents, since most present-day researchers are reluctant to grant non-human animals full-fledged metarepresentational abilities.[18]

Carruthers (2008, see also 2009) speculates about the mechanism underlying feelings of uncertainty in such cases, which he calls 'the gate-keeping mechanism': 'when confronted with conflicting plans that are too close to one another in strength [it] will refrain from acting on the one that happens to be strongest at that moment, and will initiate alternative information-gathering behaviour instead' (Carruthers 2008, p. 66). The gate-keeping mechanism operates when different goals are competing with one another to control behaviour. It initiates one of the desired behaviours only if the desires involved are not too close to one another in strength. For instance, because of the ambiguity of his visual categorizations, the subject is both weakly inclined to press the 'dense' key, and weakly inclined to press the 'sparse' key. Carruthers points out that the gate-keeping mechanism deals with the fact that 'perceptual processes are inherently noisy' (Carruthers 2008, p. 67). No two perceptual beliefs will have the same strength even given the same stimuli. Correspondingly, the subject's inclinations to act won't be stable over time, even if the world itself does not change.

Carruthers makes clear that the operations of the gate-keeping mechanism do not require metarepresentational abilities. This mechanism 'is sensitive to one *property* of desire (strength) without needing to represent that it is a *desire* that has that property' (Carruthers 2008, p. 67).

[18] See, for instance, Tomasello (1999) and Tomasello et al. (2005).

It is causally sensitive to non-intentional properties of first-order mental states, namely the strength that the subject's desires have independently of their contents.

Carruthers gives a more detailed account of the way feelings of uncertainty arise out of the operations of the gate-keeping mechanism. He suggests that they consist in 'an awareness of a distinctive profile of physiological behavioural reactions caused by the activation of the gate-keeping mechanism (including hesitating and engaging in a variety of information-seeking behaviours, such as squinting at the display or looking closer), which is experienced as aversive' (2008, p. 68). In other words, feelings of uncertainty are bodily feelings akin to aversive anxiety. They have first-order contents, insofar as they are about a kind of non-mental, bodily state.

As it stands, Carruthers' account is congenial to the Water Diviner Model and what we have said about the causal origins of noetic feelings. Feelings of uncertainty are bodily feelings that co-vary with states of uncertainty (bodily hesitations, facial tensions, etc.), as they are detected by the gate-keeping mechanism. However, his account neglects the complexity of the relationship among the gate-keeping mechanism, feelings of uncertainty, and behaviour. He seems to treat on a par all behaviours caused by states of uncertainty, whether they are of Type 1 or Type 2. His list of relevant behaviours includes 'hesitating', 'squinting at the display', 'looking closer' (Type 1), but also 'engaging in information-seeking behaviour', 'searching for another alternative' (Type 2). Obviously, 'searching for another alternative' is a highly abstract goal, which cannot be achieved by simple, pre-wired connections between states of uncertainty and behaviour. Rather, what counts as information-gathering behaviour depends on the subject's background beliefs, and hence is a highly contextualized matter.

As we have seen, the role of epistemic feelings in both types of behaviour is very different. On the one hand, implicit metacognitive processes can give rise to spontaneous simple behaviours such as pausing, squinting, moving one's head from side to side, etc. In such cases, which involve forms of procedural metacognition, conscious feelings of uncertainty are epiphenomenal; they do not *intervene* between states of uncertainty and behaviour. On the other hand, these feelings can give rise to new premises participating in further, explicit reasoning. In the latter cases, which involve forms of deliberate metacognition, feelings of uncertainty essentially intervene between states of uncertainty and more controlled behaviour.

So the situation with respect to Smith's non-human animals is more complex than Carruthers seems to suppose. There are in fact three main interpretations of Smith's results:

1. The animals have acquired a new form of procedural metacognition (a new practical skill), but they lack deliberate metacognition. If they have feelings of uncertainty, the latter are epiphenomenal and are not used in explicit practical reasoning.
2. The animals have acquired new forms of both procedural and deliberate metacognition. They can use feelings of uncertainty in explicit practical reasoning without bringing to bear metarepresentational resources (which they lack).
3. The animals have acquired new forms of both procedural and deliberate metacognition. They can use feelings of uncertainty in explicit practical reasoning as having metarepresentational contents (what they feel is that they are *unsure* about their perceptual categorizations).

What would constitute empirical evidence in favour of the animals manifesting deliberate, and not merely procedural, metacognition? Like the other types of noetic feelings, feelings of uncertainty can play an epistemic role in practical reasoning only if they can be 'at the service of many distinct projects', and their 'influence on any project [is] mediated by other beliefs', to borrow the terms used by Gareth Evans in order to characterize the distinction between explicit beliefs and implicit representations (Evans 1985, p. 337). In general, the ability to use noetic feelings as first premises in theoretical and practical reasoning requires a certain degree of *cognitive flexibility*.

Thus, the empirical hypothesis that some non-human animals can make an adaptive use of the 'uncertainty' response turns on the question of whether their behaviour has enough cognitive flexibility. In other words, the question is whether the animals' behaviour when they choose the 'uncertainty' response is spontaneous or deliberate, i.e. rationally mediated by other beliefs. This question cannot be answered just by observing a single piece of behaviour, or the same type of behaviour within a single task. Much more relevant is the finding that an animal has the ability to *transfer* (without new learning) the choice of the 'uncertainty' response across quite different tasks.[19] For this ability indicates a fair amount of cognitive flexibility, which confirms the deliberate character of the animal's response.

If, on the contrary, the animal learns to use the opt-out button but is unable to transfer its competence to other tasks, then we should say that what it acquired is merely a new procedural skill, an original piece of know-how. It knows how to use the opt-out button in a limited class of contexts, in which the same task or very similar ones are at stake. The animal's skill is still metacognitive, but only in the procedural sense. If the animal experiences noetic feelings, the latter are epiphenomenal and play no causal or epistemic role in the animal's behaviour.[20]

Assuming that the animals have acquired a genuine form of deliberate metacognition, how should we arbitrate between the second and the third interpretations? It is an open question whether cognitive flexibility, which arguably can be observed in the animal realm, requires a form of reflexivity, which some consider to be unique to humans. Of course, the kind of reflexivity that is associated with the possession of metarepresentational abilities enables a strong form of cognitive flexibility, but there may be non-reflexive forms of cognitive flexibility as well.

If room is made for the second interpretation, then Smith's results cannot be used to show that non-human animals, such as some species of monkeys, have metarepresentational abilities (and indeed Smith himself does not favour the third interpretation of his results). For these results would be compatible with the fact that noetic feelings have first-order intentional contents. However, what such contents might be has not been determined yet, and to this question I now turn.

The Competence View

In this section, I shall sketch an abstract account of the intentional contents of at least some noetic feelings, which I argue makes them first-order rather than metarepresentational. I shall call this account 'the Competence View'.

A possible strategy would be to suggest that what appears to be metarepresentational information carried by the intentional content of a noetic feeling is in fact carried at the level of its psychological *mode*. For instance, the content of the feeling of uncertainty relative to the state of affairs that *p* is not *that I feel uncertain that p*, but simply *p* itself. The relevant attitude is feeling-uncertain(*p*) rather than feeling(uncertain that *p*). My main worry with this suggestion is that it does not explain what premises feelings of uncertainty add to explicit reasoning. Of course it cannot be the premise that *p* itself. In other words, what needs to be explained is how the contents of judgements spontaneously based on noetic feelings, which correspond to the latter's acquired or 'real' contents, can fall short of being metarepresentational.

[19] See Proust (2006).

[20] Admittedly, if the concept of cognitive flexibility is vague, it will be difficult to draw the boundary between cases in which metacognition is purely procedural and cases in which it involves noetic feelings that yield first premises as a basis for reasoning to a practical conclusion.

According to the Competence View, a particular noetic feeling is about one's own cognitive competence at a given task. Its content can have the form *I can do this* (or the selfless form *This can be done*), where the demonstrative 'this' refers to a relevant cognitive task in the subject's current situation. In this respect, noetic feelings are akin to feelings of physical competence. When I walk down a rocky hill, my readiness to jump from one rock to another may be based on the feeling *that I can do it*. My feeling is about my competence in a *physical* task, namely jumping to a particular rock. What differentiates cognitive from physical tasks is a difficult question. As a first approximation, one can say that success in doing a cognitive task hangs on possessing beliefs or pieces of information that are not immediately transparent in the subject's situation. For instance, solving the bat-and-ball puzzle is a cognitive task because it requires that one *work out* the correct answer (even at the implicit level), which is not immediately given in the puzzle itself.[21]

On the Competence View, noetic feelings provide their subjects with a type of *modal* knowledge. They yield information about what might *easily* happen, now or in the near future. Something might easily happen if it is the case in nearby possible worlds (where the notion of modal proximity is context-dependent). For instance, the feeling of knowing is the feeling that one's performance is or will be successful in possible worlds close to the actual world. Now these worlds can be more or less close to the actual world, depending on the robustness of one's competence. The more robust one's competence is, the less easily one's performance might fail. If one's competence is fragile, one's performance might fail in possible worlds not too distant from the actual one. One might suggest that *degrees* of noetic feelings can then be modelled in terms of the modal extent to which one's performance is successful. A strong feeling of knowing indicates that one should not expect one's performance to fail too easily. In contrast, a weak feeling of knowing indicates that while one can still do the task, one's performance might more easily fail. In short, thanks to their noetic feelings, subjects have some information about the degree of proximity of the worlds in which their performance would succeed or fail.

The Competence View makes noetic feelings first-order *only if* one can represent one's own cognitive competence without representing it as involving beliefs or other intrinsically contentful states. The challenge is to show that the explicit target of noetic feelings is a particular task rather than the beliefs that are required to deal with it. For instance, the feeling of knowing can be the feeling that one *can* answer the question, rather than the feeling that one *knows* the answer to the question—although it is always possible (and perhaps inevitable) for adult human beings to redescribe their feelings in explicitly metarepresentational terms.[22]

However, it does not follow that all rational uses of feelings of certainty and uncertainty require metarepresentational abilities. In general, according to the Competence View, the contents of noetic feelings can be action-oriented rather than belief-oriented. They can tell the subject something about what she is doing or is inclined to do.[23] For instance, feelings of certainty in the

[21] Here is the puzzle: 'A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?' Many people answer '10 cents'. Kahneman and Frederick (2005, p. 273) comment that 'the surprisingly high rate of errors in this easy problem illustrates how lightly system 2 [the deliberate system] monitors the output of system 1 [the intuitive system]: people are often content to trust a plausible judgment that quickly comes to mind. (The correct answer, by the way, is 5 cents.)'

[22] On the uniquely human tendency to redescribe in metarepresentational terms what are in fact first-order states and processes, see Povinelli (2003). When an initially first-order state is systematically redescribed in metarepresentational terms, it may end up *acquiring* a metarepresentational content. Perhaps this is the case with feelings of knowing experienced by human adults.

[23] Then one might object that they are about one's performance rather than one's competence. Assessing one's competence is based on some concept of competence, whereas assessing one's performance is

context of a categorization task may tell the subject something like: 'If you press the 'dense' key, you are guaranteed to be successful'. In contrast, feelings of uncertainty may tell something like: 'Any success in pressing the 'dense' key will be accidental'. In a nutshell, these feelings can have contents of the form 'I can (cannot) succeed in pressing the right key'. This will be the case when what is at stake is one's success in doing a particular task rather than, more specifically, the truth of one's perceptual beliefs, even if the former actually depends on the latter.

Contents of the form 'I can do it' are not metarepresentational, at least in the sense in which contents of the form 'I believe/know that *p*' are metarepresentational. They are modal contents, which presumably entails that their grasping requires some understanding of counterfactual representations. What their grasping does not require, at least when they are used strategically in the context of practical tasks, is the ability to form representations about mental representations, i.e. to have a theory of mind.

It might be objected that even contents of the form 'I can do it' are in fact concealed metarepresentations. David Lewis notes that 'the 'can' and 'must' of ordinary language do not often express absolute ('logical' or 'metaphysical') modality. Usually they express various relative modalities' (Lewis 1983, p. 246), for instance, modalities relative to our stock of knowledge. This is also the case with the notion of competence that is expressed here by the modal verb 'can'. Noetic feelings can tell the subject something about her performance in nearby possible worlds, but what counts as a nearby world is relative to the subject's cognitive abilities, for instance the acuity of her perceptual discriminations. It does not follow, though, that noetic feelings are necessarily *about* one's cognitive abilities as such. One can be aware of a relative property without representing what the property is relative to. For instance, even if colour properties are relative to the structure of our visual system, our colour experiences do not represent our visual system as such.

Conclusion

This essay was about the psychological nature of noetic feelings. I have argued that noetic feelings are neither higher-order beliefs or memories (contra the Simple Model) nor introspective experiences of first-order epistemic states (contra the Direct Access Model). Rather, they are first-order bodily experiences, namely non-sensory affective experiences about bodily states, which given our brain architecture co-vary with first-order epistemic states, in such a way that they can be recruited, through some kind of learning or association process, to represent conditions hinging on relevant epistemic properties of one's own mind. This is what I have called 'the Water Diviner Model'.

Within this model, noetic feelings can be seen to be associated with two kinds of metacognitive abilities, which I called 'procedural' and 'deliberate'. At the procedural level, our brain realizes mechanisms whose function is to monitor the quality of our cognitive processes in order to produce spontaneous mental and/or physical behaviour (such as attempting to remember a name, reading more slowly, or moving one's head from side to side to resolve visual ambiguity). At the deliberate level, the same mechanisms can generate conscious noetic feelings, which can be further exploited in controlled reasoning to produce more context-sensitive behaviour (such as

merely based on trying to do something. However, this objection neglects the *modal* component that feelings of knowing have according to the Competence View. This is where some concept of competence (embodied in the 'can' of 'I can do it') enters the picture. Thanks to Joëlle Proust for prompting me to clarify this point.

going through the alphabet to provoke remembering, pointing to difficult sentences, or using a magnifying glass).

It follows that the question of the relationship between metacognition and metarepresentation divides into two, depending on whether procedural or deliberate metacognition is at stake. On the one hand, procedural metacognition does not require metarepresentational abilities at all, because it does not manipulate representations as of other representations. On the other hand, there is a genuine issue as to whether the (acquired) intentional contents of noetic feelings can be first-order or must be metarepresentational. One might claim that because noetic feelings track epistemic states, their contents can only be explicitly *about* them. However, the fact that subjects discriminate between knowledge and ignorance shows at best that they know *when* they know (at least sometimes), but not necessarily *that* they know. I have tentatively suggested a way of construing the contents of at least some noetic feelings, as being about one's own cognitive competence at a given task, which does not obviously tie them to metarepresentational abilities.

Of course, much more has to be said about the epistemology of noetic feelings. It is generally agreed that noetic feelings are fallible but reliable. Intuitively, though, they are not on a par with perceptual experiences, which have the property of disclosing part of the world to us. It would be odd to suggest that we can *perceive* (even amodally) our likely success in some cognitive task, in the same way that we can visually experience the presence of coffee in the cup. There may be an interesting difference between feelings of cognitive competence and feelings of *physical* competence. We are less reluctant to acknowledge that we can more or less directly perceive our own physical competence in a particular context. For instance, I can be *visually aware* that I can jump to this rock, even if (*pace* J. J. Gibson and his theory of affordances) my perception of my physical competence in this context may not be as direct as my perception of the rock itself. Nonetheless, noetic feelings merely raise the probability that their contents are true, inviting the subject to take them into account in her reasoning. They are metacognitive signals with a significant yet limited epistemic value, at least in comparison with genuine perceptual experiences. This point is no doubt connected to the fact that the contents of noetic feelings, insofar as they concern the subject's own mental and epistemic life, are acquired or derived, in contrast with the intrinsic contents of perception.

Because my interest in this essay was in the relationship between noetic feelings and metacognitive judgements, I have assumed that noetic feelings are conscious, more precisely that they have an essentially conscious aspect. Indeed, the phenomenological observation that noetic feelings belong to the 'fringe' of consciousness is congenial to Koriat's (2006) 'crossover model', according to which noetic feelings lie at the interface between implicit and explicit processes. In contrast, de Sousa (2008) suggests that feelings differ from full-fledged emotions in that they can be 'attributed at a subpersonal level'. However, perhaps there is no real disagreement here. If de Sousa suggests that metacognitive abilities can operate below the level of consciousness, I agree with him, since I have also acknowledged the existence of a procedural form of metacognition. Now de Sousa's suggestion might be interpreted as the claim that procedural metacognition involves nonconscious noetic feelings. Since I am not sure that this claim has any real explanatory bite, I am tempted to think that my disagreement with de Sousa is purely terminological. What is important is the fact that if procedural metacognition involves *conscious* feelings, the latter are epiphenomenal and do not intervene in the implicit dynamics of monitoring and control processes at the subpersonal level.

Acknowledgements

I would like to thank Joëlle Proust and the participants of her 'Metacognition' seminar at the Jean-Nicod Institute in Paris for many insightful comments on an earlier draft of this chapter.

I borrow the metaphor of noetic feelings as 'seeds' of self-knowledge from Alston's classical essay on feelings (Alston 1969).

References

Alston, W. P. (1969). Feelings. *The Philosophical Review*, 78(1), 3–34.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher and A. Koriat (Eds.) *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application*, pp. 435–59. Cambridge, MA: MIT Press.

Bjork, R. A. and Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn and R. M. Shiffrin (Eds.) *Essays in honor of William K. Estes (Vol. 1: From learning theory to connectionist theory; Vol. 2: From learning processes to cognitive processes)*, pp. 35–67. Hillsdale, NJ: Erlbaum.

Bonjour, L. (1985). *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Brown, A. S. (2000). Tip-of-the-tongue phenomena: An introductory phenomenological analysis. *Consciousness and Cognition*, 9(4), 516–37.

Brown, A. S. (2003). A review of the déjà vu experience. *Psychological Bulletin*, 129, 394–413.

Carruthers, P. (2005). *Consciousness. Essays from a Higher-Order Perspective*. Oxford: Clarendon Press.

Carruthers, P. (2006). *The architecture of the mind*. Oxford: Clarendon Press.

Carruthers, P. (2008). Meta-cognition in Animals: A Skeptical Look. *Mind and Language*, 23(1), 58–89.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32(2), 1–62.

de Sousa, R. (1987). *The Rationality of Emotion*. Cambridge, MA: MIT Press.

de Sousa, R. (2008). Inference and epistemic feelings. In G. Brun, U. Doguoglu, and D. Kuenzle (Eds.) *Epistemology and emotions*, pp. 185–204. Aldershot: Ashgate.

Dienes, Z. and Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, 22, 735–808.

Dienes, Z. and Perner, J. (2002). The metacognitive implications of the implicit-explicit distinction. In P. Chambres, M. Izaute, and P.-J. Marescaux (Eds.) *Metacognition: Process, function, and use*, pp. 171–190. Dordrecht: Kluwer Academic Publishers.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Clarendon Press.

Evans, G. (1985). Semantic theory and tacit knowledge. In G. Evans (Ed.) *Collected Papers*. Oxford: Clarendon Press.

Evans, J. St. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10), 454–9.

Evans, J. St. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, 13(4), 321–39.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–78.

Evans, J. St. B. T. and Frankish, K. (Eds.) *In two minds: Dual processes and beyond*. Oxford: Oxford University Press.

Gordon, R. (1995). Simulation without introspection or inference from me to you. In M. Davies and T. Stone (Eds.) *Mental Simulation*. Oxford: Blackwell.

Harris, P. L., Kruithof, A., Meerum Terwogt M., and Visser, T. (1981). Children's detection and awareness of textual anomaly. *Journal of Experimental Child Psychology*, 31, 212–30.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56, 208–16.

Hookway, C. (2003). Affective states and epistemic immediacy. *Metaphilosophy*, 34, 78–96.

Hookway, C. (2008). Epistemic immediacy, doubt and anxiety: on the role of affective states in epistemic evaluation. In U. Brun, U. Doguouglu, and D. Kuenzle (Eds.) *Epistemology and Emotions*, pp. 51–66. Aldershot: Ashgate.

James, W. (1879). Feelings of rationality. *Mind*, 4, 1–22.

James, W. (1980). *Principles of Psychology*. New York: Holt.

Kahneman, D. and Frederick, S. (2002). Representativeness revisited: attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.) *Heuristics and Biases: The Psychology of Intuitive Judgment*, pp. 49–81. Cambridge: Cambridge University Press.

Kahneman, D. and Frederick, S. (2005). A model of heuristic judgment. In K. Holyoak and R. G. Morrison (Eds.) *The Cambridge Handbook of Thinking and Reasoning*, pp. 267–94. Cambridge: Cambridge University Press.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General*, 124, 311–33.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126, 349–70.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149–71.

Koriat, A. (2006). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, and E. Thompson (Eds.) *Cambridge Handbook of Consciousness*, pp. 289–325. New York: Cambridge University Press.

Koriat, A. and Levy-Sadot, R. (2000). Conscious and unconscious metacognition: A rejoinder. *Consciousness and Cognition* 9, 193–202.

Koriat, A. and Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 34–53.

Koriat, A., Ma'ayan, H., and Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology*, 135(1), 36–69.

Laird, J. D. (2007). *Feelings. The Perception of Self*. Oxford: Oxford University Press.

Lewis, D. (1983). Scorekeeping in a language game. In D. Lewis (1983). *Philosophical Papers*, Vol. I, pp. 233–49. Oxford: Oxford University Press.

Loussouarn, A. (2010). De la métaperception à l'agir perceptif. PhD thesis, EHESS, Institut Jean-Nicod.

Mangan, B. 2001. Sensation's ghost. The non-sensory 'fringe' of consciousness. *PSYCHE*, 7(18).

Matthen, M. (2005). *Seeing, Doing, and Knowing*. Oxford: Oxford University Press.

Metcalfe, J. (2000). Feelings and judgments of knowing: Is there a special noetic state? *Consciousness and Cognition* 9, 178–86.

Metcalfe, J. and Shimamura, A. P. (Eds.) *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press.

Nelson, T. O. (Ed.) (1992). *Metacognition. Core Readings*. Cambridge, MA: MIT Press.

Povinelli, D. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Science*, 7(4), 157–60.

Price, M. C. and Norman, E. (2008). Intuitive decisions on the fringes of consciousness: Are they conscious and does it matter? *Judgment and Decision Making*, 3(1), 28–41.

Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of Emotion*. New York: Oxford University Press.

Prinz, J. (2007). *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Proust, J. (2006). Rationality and metacognition in non-human animals. In S. Hurley and M. Nudds (Eds.) *Rational Animals*, pp. 247–74. Oxford: Oxford University Press.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 159, 271–95.

Proust, J. (2008). Epistemic agency and metacognition: An externalist view. *Proceedings of the Aristotelian Society*, 108, 241–68.

Reber, R., Schwarz, N. and Winkielman P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8(4), 364–82.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology*, 19, 90–138.

Reder, L. M. and Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 435–51.

Reder, L. M. and Shunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.) *Implicit memory and metacognition*, pp. 45–78. Hillsdale, NJ: Erlbaum.

Rensink, R. A. (2004). Visual sensing without seeing. *Psychological Science* 15(1), 27–32.

Rosenthal, D. (2000). Consciousness, content, and metacognitive judgments. *Consciousness and Cognition*, 9, 203–14.

Schwarz, B. L. (2002). *Tip-of-the-tongue states: Phenomenology, mechanism, and lexical retrieval*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwarz, N. and Clore, G. L. (1996). Feelings and phenomenal experiences. In E. T. Higgins and A. W. Kruglanski (Eds.) *Social psychology: Handbook of basic principles*, pp. 433–65. NewYork: Guilford Press.

Smith, J. D. (2005). Studies of uncertainty monitoring and metacognition in animals and humans. In H. S. Terrace and J. Metcalfe (Eds.) *The Missing Link in Cognition. Origins of Self-Reflective Consciousness*, pp. 242–71. Oxford: Oxford University Press.

Smith, J. D. (2009). The study of animal metacognition. *Trends in Cognitive Sciences*, 13(9), 389–96.

Smith, J. D., Shields, W. E., and Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26, 317–73.

Spehn, M. K. and Reder, L. M. (2000). The unconscious feeling of knowing: a commentary on Koriat's paper. *Consciousness and Cognition*, 9, 187–92.

Stanovich, K. E. and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645–726.

Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard, MA: Harvard University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28, 675–735.

Thomson, V. (2008). Dual-process theories. A metacognitive perspective. In J. St. B. T. Evans and K. Frankish (Eds.) *In two minds: Dual processes and beyond*, pp. 171–195. Oxford: Oxford University Press.

Tye, M. (2009). *Consciousness revisited*. Cambridge, MA: MIT Press.

Whittlesea, B. W. A. and Williams, L. D. (2001a). The Discrepancy-Attribution Hypothesis: I. The Heuristic Basis of Feelings of Familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1), 3–13.

Whittlesea, B. W. A. and Williams, L. D. (2001b). The Discrepancy-Attribution Hypothesis: II. Expectation, uncertainty, surprise, and feelings of familiarity. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 27(1), 14–33.

Winkielman, P. and Fazendeiro, T. (in preparation). Effects of conceptual fluency on memory and liking.

Winman, A. and Juslin, P. (2005). 'I'm m/n confident that I'm correct': Confidence in foresight and hindsight as a sampling probability. In K. Fiedler and P. Juslin (Eds.) *Information sampling and adaptive cognition*, pp. 409–39. Cambridge: Cambridge University Press.

Chapter 20

Metacognitive perspectives on unawareness and uncertainty

Paul Egré and Denis Bonnay

Introduction

What does it take to know that one does not know something, or to know that one knows? In epistemic logic, the ability to know that one does not know something whenever such is the case is referred to as 'negative introspection', and distinguished from the property of 'positive introspection', which concerns the ability to know that one knows whenever one knows (see Fagin et al. 1995).

Negative introspection is usually seen as a very demanding condition on knowledge, for two reasons: one reason, which will be examined more carefully later, is that we often fail to know that we don't know p simply because we do not have access to the basic ingredients of the proposition p. Another reason is that systematic knowledge that one does not know would prevent us from believing that one knows in all cases in which we have false beliefs: suppose that I falsely believe p, and therefore that I fail to know that p; in order to know that I don't know p for a fact, I would need to believe that I don't know p, and so I would have to believe both that I know p and that I don't know p, a contradiction.

There is a more active debate in the philosophy of knowledge regarding whether positive introspection should be seen as an analytic property of our knowledge. On the Cartesian view of knowledge, to know a proposition normatively requires us to be able to know that we know. This view was defended by Hintikka in particular and is considered as characteristic of epistemological internalism. Hintikka's argument rests on the idea that if knowing implies having a conclusive justification for one's belief, then one should thereby be in a position to know that the justification is conclusive (see Hintikka (1970) and Hemp (2006) for an overview). One criticism of Hintikka's analytic conception of the relation between knowing and knowing that one knows has been that knowing that one knows would require one to possess the concept of knowledge. This would deny knowledge to animals and infants, who presumably do not have the concept of knowledge. This, however, presupposes that knowing that one knows necessarily involves a metarepresentational attitude towards one first-order knowledge, an assumption that is far from obvious in the light of the idea that metacognitive processes need not always involve an abstract representation of the concept of knowledge (see Proust (2007) for a discussion). Another externalist attack on the analytic connection between knowing and knowing that one knows comes from considerations about the structure of justification. If to know a proposition is to have a good justification for it, then one may know without always knowing that one knows, for otherwise the risk is that of an infinite regress in the justifications. This, in a nutshell, is the gist of epistemological externalism about knowledge.

In this paper we shall not focus on normative considerations about the well-foundedness of either positive or negative introspection, in part because we contributed to that debate elsewhere

(see, in particular, Egré 2008; Bonnay and Egré 2009, 2011). Rather, the aim of this chapter will be to discuss some of the constraints on one's abilities to know that one knows and to know that one does not know in relation both to epistemology and to psychology. Obviously, there are many ways in which we can fail to realize our ignorance. On the other hand, there are also situations in which we have a clear access to our knowledge as well as to our ignorance. What makes the difference between those? Part of the present contribution will be an effort to interpret actual psychological data and relate them to more abstract models of knowledge and uncertainty (as used in epistemic logic; see Egré and Bonnay (2010) for an example, in this chapter, we shall remain deliberately informal as far as possible).

Our leading thread in this discussion will be the distinction made in formal epistemology between two forms of ignorance, namely uncertainty and unawareness. Fundamentally, unawareness can be defined as a form of ignorance resulting from the lack of conceptual or representational resources needed to articulate a proposition. Uncertainty, on the other hand, concerns the lack of evidence needed to adjudicate the truth or falsity of a proposition that one can represent and articulate. Uncertainty, most of the time, is conscious, while unawareness, by definition, is unconscious. Our proposal is to examine the implications of the distinction between uncertainty and unawareness for metacognition. What we shall argue is that knowing that one knows or that one does not know is typically harder and less reliable in situations that require us to evaluate the strength of one's uncertainty. In contrast to that, knowing that one does not know is easier and more reliable for unknowns grounded in antecedent unawareness.

The chapter is structured as follows. In the second section ('Uncertainty and unawareness'), we introduce the distinction between uncertainty and unawareness and review the main differences between them. In the third and fourth sections ('Unconscious ignorance and implicit knowledge' and 'Underconfidence and overconfidence'), we use the distinction to classify different ways in which one may fail to know that one knows or that one does not know a proposition. In the fifth section ('Evaluating one's ignorance'), we proceed to the discussion of two sets of experimental data concerning metacognition: experiments by Glucksberg and McCloskey (1981), and more recently by Hampton and colleagues (2012), concerning the evaluation of one's ignorance, and experiments by Smith et al. (2003) concerning the monitoring of one's uncertainty. Consistent with the model of Glucksberg and McCloskey, we shall argue that appreciating the strength of one's evidence and appreciating the availability of specific conceptual resources in memory likely involve different mechanisms. More generally, we argue that the distinction between uncertainty-based unknowns and unawareness-based unknowns can be subsumed under Glucksberg and McCloskey's two-stage model for decisions about ignorance. In the sixth section ('Evaluating one's uncertainty'), finally, we focus on higher-order knowledge about one's uncertainty and discuss the case for an asymmetry between knowing that one knows and knowing that one does not know.

Uncertainty and unawareness

A fruitful way to approach the definition of metacognitive abilities such as knowing that one knows or that one does not know is to start by an examination of the notion of ignorance. The object of this section is to argue that ignorance, understood as the failure to know some proposition, results from two importantly distinct sources, which are called *uncertainty* and *unawareness* in the epistemological literature.

Two forms of ignorance

While the notion of uncertainty has been at the centre of epistemic logic since its inception, the clarification of the concept of unawareness is much more recent (see, in particular, Franke and de

Jager (2010) for an excellent exposition and overview). As we shall see, uncertainty and unawareness themselves come in different varieties. However, the main opposition we can draw between them concerns the extent to which either form of ignorance is accessible to consciousness.

Let us consider uncertainty first. Suppose I am playing a version of the Monty Hall game.[1] I am faced with two doors labelled A and B. Behind one of those, there is a goat, and behind the other there is a car. To win the game is to open the door with a car behind it. The quizmaster informs me of the situation and asks me which door I want to open. Clearly, I am in a state in which I do not know whether the car is behind door A or behind door B. In this case my ignorance about whether the car is behind door A or behind door B results from my incapacity to discriminate between the two doors. I entertain two possibilities about the true state of the world: that the car is behind door A and that the car is behind door B. The fact that these two possibilities are equally open to me is what we call *uncertainty*.

Contrast the previous situation with the following. Suppose I never heard of J. R. R. Tolkien, the author of *The Hobbit*, nor had I come across any of his books. In that situation I am ignorant of a number of facts about Tolkien besides his existence. In particular, I fail to know that Tolkien is the author of *The Hobbit*. My ignorance in that case is quite different from my ignorance in the previous case. The situation is not one in which I am uncertain as to who the author of *The Hobbit* might be. In particular, it is quite different from a situation in which I might have read *The Hobbit* and come across the name of Tolkien several times before, but in which, asked about who the author is, I would hesitate between J. R. R. Tolkien and C. S. Lewis. In the former situation, as opposed to the latter, I do not have the wherewithal to even represent the proposition that Tolkien is the author of *The Hobbit*. In a case like this I am simply *unaware* that Tolkien wrote *The Hobbit*, because I do not have the ingredients needed to entertain that proposition.

In the words of Heifetz et al. (2006), 'unawareness refers to lack of conception rather than to lack of information'. *Lack of conception* corresponds to a state in which we cannot verbally or conceptually articulate a possibility. Lack of conception can mean different things, however. The most radical form is what we may call *lack of acquaintance* with the concepts, when we do not even have the ingredients available in memory to articulate the proposition. This corresponds to the example we just discussed. In a lot of cases, however, lack of conception can result from *inattentiveness*, as discussed by Franke and de Jager, namely when the conceptual resources are available in memory, but when we are temporarily blind to them. Franke and de Jager, for instance, give the example of someone looking for his keys, and temporarily failing to even represent the possibility that the keys might be in his car. On their account, this is not lack of acquaintance with the car or the concept of the car, but temporary blindness to the possibility that the keys might be in the car, due to a temporary failure to activate the representation of the car in one's memory. As Franke and de Jager put it, the subject then would be able to utter truths like: 'the keys are either on the desk or not on the desk', but would not be able to say in the same way: 'the keys are in the car or not in the car'.

This form of inattentiveness, finally, should be distinguished from a more common form of so-called unawareness, for cases in which we do have the conceptual resources available in memory, can activate them, but simply discard them as irrelevant. Such cases are certainly typical of most of our false beliefs. For a long time, for instance, I used to assume that teaine and caffeine were two chemically different substances. Later, I was told that they were the same molecule, and came to revise my earlier belief. So it could be said that I was *unaware* that teaine and caffeine were the same substance until I was told, not because I could not conceptually articulate the possibility that they were the same substance, but because that was a possibility I was failing to

[1] Wherever we refer to first-person experience, we deliberately use the pronoun "I" in what follows.

entertain as open. In the words of Franke and de Jager, I was *assuming* that teaine and caffeine were different substances and this blind assumption could be construed as some form of unawareness.

However, the situation is very different from those involving the two kinds of unawareness previously discussed. Lack of conception is symmetric. If I cannot represent the proposition that Tolkien is the author of *The Hobbit*, I cannot represent either the proposition that Tolkien is not the author of *The Hobbit*. Similarly, if I overlook the possibility that the keys might be in the car, I do not consider the possibility that the keys might not be in the car as a salient possibility (that is to say, I will not be searching places *qua* places that are not places in the car). By contrast, the teaine versus caffeine example is asymmetric. I am overlooking the possibility that they are same but I am precisely not overlooking the possibility that they are different. From now on, we shall reserve the label 'unawareness' to symmetric cases, which stem from lack of conception or from inattentiveness.

Main differences

The difference between uncertainty and unawareness can now be characterized more abstractly. In ordinary ascriptions of knowledge, first of all, note that we say of someone that they do not know *whether p*, or that they are uncertain about *whether p*. By contrast, we report situations of unawareness by saying of someone that they do not know *that p*, or that they are unaware *that p* is true. To fail to know whether or not *p* is to entertain *p* as well as its negation as two open possibilities. By contrast, cases of unawareness are cases in which the fact that *p* is not entertained as a possibility, and in which the contradictory alternative, consequently, is not even represented in the agent's mind.[2]

Secondly, uncertainty and unawareness are resolved in different ways. Uncertainty reduces, and knowledge increases, as possibilities gradually get eliminated. Consider the Monty Hall game again, and suppose that I randomly pick out door A. The quizmaster opens it; unfortunately I see a goat behind it. Given my new evidence, however, I now eliminate the possibility that the car is behind door A, from which I can infer that it is behind door B. An increase in awareness, by contrast, is not adequately pictured as the narrowing down of a set of epistemic possibilities. Intuitively, it corresponds to the opposite. In a case in which I already have the conceptual resources but fail to attend to *p* as a possibility, becoming aware means expanding the set of possibilities initially thought to be relevant. For instance, my becoming aware that teaine and caffeine were the same substance implied the consideration of a possibility previously excluded by my belief. In cases in which one is unaware of a proposition due to lack of the conceptual resources necessary to articulate the proposition, becoming aware will not quite mean expanding the set of possibilities. Rather, it involves adding structure to the space of possibilities. For instance, if I had never heard of Tolkien nor of the novel *The Hobbit*, and come across both names, I acquire the capacity to ask a new question such as: 'Is Tolkien the author of *The Hobbit*?'. I can thereby divide the space of logical possibilities by means of a division that was previously unavailable to me (see Bromberger (1987/1992) on what it takes to articulate one's ignorance of a proposition, and Pérez Carballo (submitted) for a recent account generalizing on that idea; we refer to the next section for an example).

The third and most relevant difference between unawareness and uncertainty, finally, concerns the status of consciousness in relation to epistemic possibilities. In a state of uncertainty, an agent

[2] For more on the distinction between 'knowing that' and 'knowing whether' constructions, see Aloni et al. in press.

is consciously entertaining possibilities as open, and consciously trying to get more information in order to reduce that uncertainty. In a state of unawareness, by definition, the agent cannot be conscious of the possibilities he or she is failing to take into account. The distinction between conscious and unconscious possibilities has a metacognitive import. As Franke and de Jager put it, unlike uncertainty, 'unawareness is not introspective' (see Dekel et al. (1998) for a formal account of unawareness based on this important observation). This means that whereas one can know that one experiences uncertainty simultaneously with that uncertainty, one cannot know that one is unaware of a proposition at the moment one is unaware of it.[3] Knowing that one knows and knowing that one does not know are thus likely to obey different constraints depending on whether one's ignorance is a matter of uncertainty or of unawareness.

Unconscious ignorance and implicit knowledge

This section reviews some of the ways in which a state of unawareness precludes knowing that one does not know, but also knowing that one knows. In both cases, becoming aware implies a transition from implicit to explicit uncertainty, or relatedly, from implicit to explicit knowledge.

Unconscious ignorance

A situation in which an agent lacks the basic concepts necessary to articulate a proposition will necessarily prevent her from being conscious that she does not know a proposition at the moment she does not know it. Someone who never heard of Tolkien and *The Hobbit* cannot know that she does not know that Tolkien is the author of *The Hobbit*. However, if asked the question: 'Who is the author of *The Hobbit*?', or 'Is Tolkien the author of *The Hobbit*?', the mention of these names can be sufficient to trigger a change of state in the agent, from unawareness to uncertainty. This means that, at the moment the agent is asked the question, the agent's unawareness disappears, and the agent is now in a position to ask all sorts of questions about Tolkien. To the extent that the agent accommodates the information that expressions like 'Tolkien' or 'The Hobbit' have a reference, and that they belong to the expected referential categories ('Tolkien' refers to a person, 'The Hobbit' to some work of art) the agent is ipso facto in a position to know that she does not know whether or not Tolkien is the author of *The Hobbit*.

A convenient way of picturing the transition from unconscious to conscious ignorance is as follows. Let p stand for the sentence 'Tolkien is the author of *The Hobbit*', and $\neg p$ for its negation. Unawareness can be represented by the agent's incapacity to *delineate* between p and non-p possibilities. In the left part of Fig. 20.1, the rectangle represents the space of conceptual possibilities, divided between p and $\neg p$ as distinct possibilities. The absence of conscious delineation is represented by a dashed line between p and non-p regions. Uncertainty, on the other hand, results from the agent's limited capacity to *discriminate* between possibilities. On both figures, this uncertainty is represented by the fact that the agent's information state, the set of possibilities available in principle to the agent on the basis of her evidence, overlaps on p and $\neg p$ regions. In the left-hand diagram, uncertainty is only implicit, however. The transition from implicit to explicit uncertainty corresponds to the replacement of a dotted line by a solid line between p and non-p possibilities. The agent's information state is such that, once the agent gains awareness of

[3] We do not rule out the possibility of states of unconscious uncertainty, but the point is that being uncertain is compatible with the consciousness of that state, whereas a state of unawareness is incompatible with the simultaneous consciousness of one's unawareness. In other words, I can be conscious that I *am* uncertain, whereas I can only realize that I *was* unaware. For more on what it means to dynamically realize that one was ignorant, see van Benthem (2004) and Bonnay and Egré (2011).

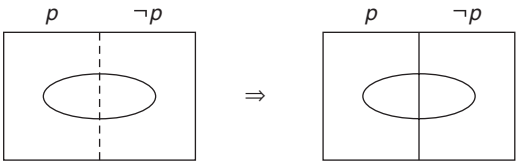


Fig. 20.1 From implicit to explicit uncertainty.

the proposition at issue ('is Tolkien the author of *The Hobbit*?'), she is in a state of conscious uncertainty about the answer to that question.

Implicit knowledge

From Fig. 20.1 we can account for dual cases, in which the agent lacks the conceptual resources to even figure out a particular question, but such that, if she were given the appropriate concepts, she would correctly discriminate between right and wrong answers. One speaks of *tacit* or *implicit* knowledge for such cases. A good example of such tacit knowledge is linguistic knowledge. In the tradition of generative grammar, most of our linguistic competence is characterized as a form of tacit knowledge of regularities about sentence formation. Halle in particular describes phonological knowledge as 'knowledge untaught and unlearned' (Halle 1978). To illustrate it, he gives the example of phonological rules that competent speakers of English master without difficulty, but are unaware of, such as plural formation in English. There are three kinds of morphophonological realization of the plural in English, namely plurals in -iz- as in *buses*, plurals in -s- as in *cats*, and plurals in -z- as in *dogs*. A regularity about the choice between these plural suffixes is, for instance, that: 'If a noun ends with a sound that is non-voiced, the plural is formed with -s-'. For instance, the noun 'dog' ends with a voiced stop, whereas the noun 'cat' ends with a non-voiced stop. The knowledge of such a generalization is obviously implicit in most speakers, simply because the concepts of a stop or of a voiced consonant need not be available.

If an agent were given those concepts, she may still not be able to thereby state the generalization of course. However, consider a simpler instance of this generalization. There is obviously a sense in which every competent speaker of English knows that the plural of 'cat' is 'cats' rather than 'catz' or 'catiz'. Most speakers of English will be unaware of this proposition, simply for failing to attend to the possibility that the plural of 'cat' might have been 'catz' or 'catiz'. When asked, however, they would obviously respond correctly to the question: 'Is the plural of "cat" "cats", "catz", or "catiz"?'. There is a sense, therefore, in which competent speakers of English know implicitly that the plural of 'cat' is not 'catz', but are unaware of this fact, and are unaware that they know it. When asked explicitly, however, they are put in a position to correctly eliminate the possibility that the plural of 'cat' is 'catz'.[4] In Fig. 20.2, the diagram on the left represents a situation in which the agent's informational state implies such a proposition p, but in which the agent cannot initially delineate between p and $\neg p$ possibilities. Once the delineation is made, however, the informational state of the agent puts them in a position to explicitly know that p, and also to know that they know.

The example of linguistic knowledge given here to illustrate this transition from implicit to explicit knowledge may still raise some questions. It may be argued that linguistic competence is

[4] See Schaffer (2008) for a discussion of the epistemological implications of contrastive knowledge attributions of the form 'knowing that p rather than q'. See Aloni and Egré (2010) for a discussion of Schaffer's view on epistemological contrastivism.

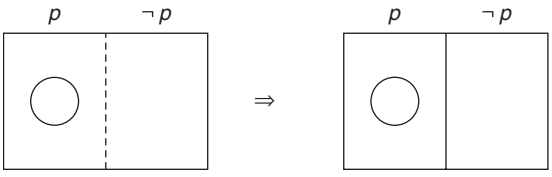


Fig. 20.2 From implicit to explicit knowledge.

a form of knowing-how, irreducible to the knowledge of a body of propositions. This view is not uncontroversial (see, in particular, Stanley and Williamson (2001) for a defence of knowing-how as propositional knowledge). Even if we grant it, however, we may still be able to find examples of implicit propositional knowledge that fit exactly the sort of transition intended by Fig. 20.2. Maybe the oldest example is Plato's discussion, in *Meno*, of what it takes for a young child to know that the length of the diagonal of the square is not commensurable with the length of the sides. Plato's argument, in a nutshell, is that the child could not come to the knowledge of that proposition if he did not know it previously. One way to make sense of Plato's views would be in the terms of Fig. 20.2: the child has all the discriminative resources to recognize that that proposition is true. What the child misses, however, are the concepts and intermediate constructions that allow him to delineate logical space in a way that will allow him to identify the relevant state of affairs (in the dialogue, Socrates teaches the concepts and constructions to the young child). A recent account of mathematical knowledge along exactly those lines is elaborated in the work of Pérez Carballo (submitted), based on the discussion of other actual examples of mathematical discoveries.

Underconfidence and overconfidence

Cases of unawareness of the kind discussed in the previous section, in which we fail to realize what we know and don't know, cannot be characterized as cases of *misrepresentation* of one's knowledge, but rather, as cases of *unrepresentation* of the structural components of a proposition. In this section we move to cases in which negative and positive introspection fail not because of unawareness, but because of a misrepresentation of the structure of one's first-order uncertainty. This corresponds to cases in which an agent has a wrong appreciation of her discrimination capacities, or of what her evidence allows her to conclude. Situations of that kind are generally described as cases of underconfidence or of overconfidence.

Overconfidence

Consider overconfidence first. Many are the occurrences in which an agent holds an incorrect belief about whether p, yet represents herself as knowing p. For instance, at the time I used to believe that teaine and caffeine were different molecules, my knowledge of chemistry did not actually allow me to rule out that they were the same. Yet I represented my evidence as conclusive enough to rule out this possibility. The situation is depicted on the left-hand diagram of Fig. 20.3. The p area represents the proposition that teaine and caffeine are different, and $\neg p$ the proposition that they are the same. The circle on that figure represents the possibilities among which I am able to discriminate on the basis of my actual knowledge of chemistry. The ellipse, on the other hand, gives the representation I have of my evidence. In this case, I believe that I know p, because the ellipse is included in the set of p possibilities. This represents a case in which I have a wrong appreciation of my actual evidence: I underestimate some possibilities, and overestimate others.

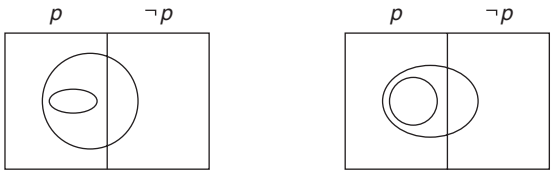


Fig. 20.3 Overconfidence and underconfidence: circle = actual first-order uncertainty; ellipse = representation of one's first-order uncertainty.

Incidentally, this diagram also represents one form of unawareness, distinct from what we characterized as 'lack of conception'. That is, it represents a situation in which I am able to articulate the difference between p and $\neg p$ possibilities and have them in my conceptual apparatus (as materialized by the solid line between p and $\neg p$), but in which I mistakenly fail to entertain $\neg p$ possibilities as open.

Underconfidence

Cases of underconfidence are exactly symmetric. In a situation of underconfidence, I am in a position to know p, but I don't adequately represent myself as knowing p. This is represented on the right-hand diagram of Fig. 20.3. This would be a situation in which I am in a position to exclude that teaine and caffeine are different substances, but in which I misrepresent my actual evidence, and believe that I don't know whether teaine and caffeine are the same or not. From a psychological point of view, underconfidence may have many different sources. From an epistemological point of view, it implies considering irrelevant possibilities as relevant.

Note that agents with a perfect capacity to evaluate their actual first-order uncertainty would be agents for whom the regions delimited by circle and ellipse would exactly coincide. By way of consequence, the lack of coincidence here between the two levels gives counterexamples to the principles of positive and negative introspection. On the left-hand diagram in Fig. 20.3, the agent does not know p, but believes she knows p. If our agent is consistent, this implies that she does not know that she does not know (for knowing that one does not know would imply believing that one does not know). On the right-hand diagram, the agent knows p, but believes she does not know p. This implies that she does not know that she knows in at least one sense, the sense in which knowing that one knows would imply believing that one knows.[5]

Inclusion or even overlap between ellipse and circle as in our figure may not always happen, except to represent specific properties relating first-order knowledge and the representation of one's knowledge. For instance an agent may completely misrepresent the evidence underlying her first-order knowledge, and believe that she knows $\neg p$ when she is actually in a position to know p (See Fig. 20.4). Intuitively, this would be a case of delusion, rather than of underconfidence, namely a case in which the agent's representation of her evidence is entirely dissociated from what one might consider as conclusive evidence (see, for instance, Feinberg and Roane (2005) for a review of clinical cases, in which an agent, for example, reports that her arm or leg, which she cannot move, belongs to someone else. This is a case in which, arguably, the agent believes her

[5] Importantly, however, in order to correctly believe that one knows p or that one does not know p on a particular occasion, a perfect representation of one's first-order uncertainty is not needed. What suffices for adequately believing that one knows p or that one does not know that p is for the two levels to draw the same distinctions with regard to p and $\neg p$ possibilities. This means that the agent may still misrepresent her knowledge about other propositions, but in a way that may be irrelevant for what concerns p.

arm to belong to someone else, and believes she knows it, where she should rationally conclude that the arm really is her arm).[6]

Another case of dissociation between first-order evidence and the representation of one's first-order evidence that might be used to illustrate such dissociations is blindsight. In blindsight, the agent who suffers brain damage has very good reasons to believe that he won't navigate a course of obstacles, but his behaviour shows he can navigate (see the overview of de Gelder 2010). The agent's internal evidence, in this case, puts him in a position to believe he cannot navigate. His behaviour, on the other hand, leads one to conclude that he has the ability to navigate. Note that blindsight is a case of dissociation between a practical ability (what one can do) and the representation of that practical ability (what one believes one can do). Whether practical abilities can be analysed in terms of propositional knowledge is a much debated issue (see Ryle (1971) and Stanley and Williamson (2001) for opposing views, as well as Lihoreau (2008)). If we accept this reduction, then blindsight can be described as a case of dissociation between knowing-how and second-order knowledge of that knowing-how. A difficult question that we shall not attempt to discuss any further here concerns the relation between what we call first-order evidence and the representation of that first-order evidence. Obviously, in blindsight the internal evidence of blindness is distinct from the external evidence and feedback that one can reliably detect obstacles, but both can eventually be accommodated by the patient. Possibly, agents with blindsight who are told that they can navigate a course of obstacles are able to rely on that further evidence to correct their initial belief that they cannot navigate. The case, in this regard, is very distinct from cases of delusion in which patients are resilient against accepting new evidence against their initial belief.

A caveat is in order regarding the way the examples we reviewed were described. For these various cases of underconfidence and of delusion to count as counterexamples to positive introspection, they need to be so analysed that the agent is taken to know but not to know that she knows. But the ascription of first-order knowledge is debatable. In the asomatognosia case and in the blindsight case, one might be tempted to resist the ascription of first-order knowledge and rather say that the agent possesses the relevant first-order evidence but somehow fails to gain first-order knowledge on the basis of that evidence. A similar analysis of the modified teaine versus caffeine example, where my knowledge in chemistry is in principle sufficient for me to exclude the possibility that they are different, could be proposed. According to that alternative analysis, I would be in a position to know that teaine and caffeine are the same but I would not thereby know that they are the same. The underlying claim would be that being in position to know does not amount to knowing. This line of thought would of course be welcome to the advocates of positive introspection who hold that positive introspection cannot fail to hold because it is a characteristic property of knowledge.

Evaluating one's ignorance

In this section we turn to the discussion of the metacognitive mechanisms by which we appreciate whether we know that we know or know that we don't know a proposition. Based on the main

[6] Agents whose arm of leg is paralysed due to hemiplegia have some internal evidence that the arm might not belong to them, for instance, inasmuch as they cannot move it. However, while some patients are ready to take other evidence into account to the effect that the limb still is theirs, others persistently deny the evidence to that effect. See Feinberg and Roane (2005, p. 667): 'Patients who have asomatognosia may attribute ownership of the limb to the examining doctor. This simple misattribution often can be reversed when the error is demonstrated to the patient. In other patients, the misidentifications are truly delusional, and patients maintain a fixed belief in the misidentifications when they are confronted with evidence of their errors.'

distinction between conceptual unawareness and informational uncertainty, we will argue that we should distinguish two sets of metacognitive abilities. One concerns the ability to monitor one's uncertainty, the other concerns the ability to appreciate one's acquaintance with the ingredients of a proposition. The section is organized as follows: first, we review in a bit more detail aspects of the modelling of knowledge in epistemic logic. Then, we consider experiments done by Glucksberg and McCloskey, and more recently by Hampton and colleagues, suggesting that the distinction between two forms of ignorance has a metacognitive correlate.

Informational content and conceptual content

Ignorance, we said so far, can result from two different conditions: lack of discrimination between competing alternatives on the one hand, lack of conceptualization of the alternatives on the other. In epistemic logic, standard models of knowledge generally incorporate the notion of discrimination. Relative to a context w, the information available to agent i is represented by a set $R_i(w)$ of possibilities compatible with i's evidence, called the information set. The information set is what we represented by a circle in Fig. 20.4, and it represents the possibilities among which the agent cannot discriminate. The agent i is then said to know p in the context w if and only if his or her information set $R_i(w)$ entails the proposition expressed by p.

Note that on that view, a knowledge state is defined purely in terms of the objective information available to the agent. A main limitation of this approach of knowledge is that it makes knowledge coarse-grained: two sentences with logically equivalent informational contents are such that an agent who knows the one is predicted to know the other. Likewise, an agent who knows S is predicted to know S' whenever the informational content of S entails that of S'.[7] For instance, the model predicts that someone who knows that the keys are in the house thereby knows that the keys are in the kitchen or in some other room. Arguably, however, one may be aware that the keys are in the house without being aware that they are in the kitchen or in some other room, because the keys being in the kitchen is not a salient possibility.

Models of unawareness offer to solve this problem by enriching the purely informational definition of knowledge. On the resulting view, knowledge is relative not only to the agent's informational set $R(w)$, namely to a set of epistemic alternatives, but it is also relative to an awareness set $A(w)$, consisting of the concepts the agent is aware of. For instance, in order to be aware that the keys are in the car or not in the car, the concept of a 'car' must be present in the agent's awareness set (see Franke and de Jager 2010). A knowledge content, on that perspective, is no longer defined in purely informational terms, but has a syntactic or linguistic structure (viz. Fagin and Halpern 1988; Franke and de Jager 2010; Hill 2010; Cozic 2011).

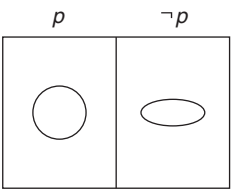


Fig. 20.4 A situation of delusion: the agent believes she knows *not p* (ellipse); her first-order evidence (circle) should lead her to conclude *p*.

[7] See Stalnaker (1990) for a discussion of this prediction, and Fagin et al. (1995) for various ways of dealing with this problem, known as the problem of logical omniscience in epistemic logic.

From a psychological point of view, therefore, the upshot of those models is that knowledge is a function of two distinct parameters: one concerns the occurrence in memory of the concepts necessary to articulate a proposition. The other concerns the quality of the information or evidence received to adjudicate whether that proposition is true or not. A reasonable prediction, therefore, is that metacognition, understood as the ability to know whether we know or don't know a proposition, will in turn be constrained differently, depending on whether we need to check on informational or conceptual content regarding our first-order knowledge.

Glucksberg and McCloskey's two-stage model

Empirical support for the hypothesis formulated here can be found in work by Glucksberg and McCloskey (1981), and more recently by Hampton et al. (2012) on the psychology of 'known unknowns'. Glucksberg and McCloskey put forward the supposition that there are two types of don't know decisions. The first type, on their account, concerns cases in which 'no potentially relevant information is known' (1981, p. 312). The second type, by contrast, concerns cases in which 'some potentially relevant information is retrieved, but this information turns out to be insufficient to permit an answer to the question'.

An imaginary example of a don't know of the first kind they give is:

1. What is the name of the largest store in Budapest?

For such a question, they consider that 'most people [not living in Hungary] would probably say something like, 'I have absolutely no idea', or 'I know nothing about *any* of the stores in Budapest'. (ibid.). To illustrate a don't know of the second kind, they give the example of the following question:

2. Is Kiev in the Ukraine?

This time, they imagine that the subject knows Kiev to be a city in Russia, knows Ukraine to be in Russia,[8] but is 'just not sure whether or not Kiev is in the Ukraine'.

Both examples are examples in which the agent, once asked the question, is in a state of ignorance about the answer. In the first case, however, the agent lacks the conceptual information needed to answer the question (a wh-question, or constituent question). In the second case, by contrast, the agent simply lacks conclusive evidence to ascertain a yes vs. no answer to the question (a polar question, or yes–no question). An intuitive difference between the two cases is that, in the former case, the agent can be confident that he or she does not know the answer. In the latter case, by contrast, the agent should be less confident that she doesn't know the answer.

The difference between the two situations points to a metacognitive difference concerning the reliability of our don't know judgements. Thus, what Glucksberg and McCloskey predicted and tested for was that unknowns of the first kind should give rise to 'rapid don't know responses', and unknowns of the second kind to comparatively slow responses. A rapid don't know answer, on their account, is the output of a preliminary phase in which relevant information is searched for in memory. If no item is found, a don't know answer is produced. A slow don't know answer, by contrast, is the output of two phases: in the first memory search phase, some relevant information is found in memory about the ingredients of the question; in the second phase, evidence is examined to establish the relation between those ingredients.

In a first experiment, Glucksberg and McCloskey presented subjects with 24 sentences such as 'John has a pencil', 'Bill doesn't have a magazine'. They first trained subjects to remember the content of the sentences to a reliable extent. In the test phase, subjects were then presented with

[8] It is debatable, to say the least, whether in 1981 'Ukraine is in Russia' was true, as assumed by Glucksberg and McCloskey. Fortunately, this does not matter to the present discussion.

sentences of six different types (true affirmative/negative, false affirmative/negative, don't know affirmative/negative) and given the choice to respond by True, False, or Don't Know. In this setting, a sentence such as 'John doesn't have a pencil', for instance, is a false negative, because 'John has a pencil' was among the 24 sentences presented. 'John has a magazine', on the other hand, is a don't know affirmative, assuming that neither that sentence, nor its negation, were initially among the 24 sentences presented.

According to the two-stage model, correct 'I don't know' responses are the output of a process terminating earlier. Correctly responding 'I don't know' only involves going through the first phase of memory search (no evidence is found, the process terminates). By contrast, correctly responding 'Yes' or 'No' involves going through the first phase of memory search (evidence is found, which needs to be examined) and also through the second phase during which evidence is examined. Therefore, Glucksberg and McCloskey's model predicts that response times for correct 'I don't know' should be faster. Indeed, Glucksberg and McCloskey found response times for correct Don't Know answers to be significantly shorter than response times for correct True and correct False.

What happens with incorrect don't Knows in this experiment? Glucksberg and McCloskey are silent about this. What the model predicts is indeed not so clear. It could be that no evidence has been found. In this case, the process would have terminated at the end of the first phase and shorter response times are to be expected. Or it could be that some evidence has been found, which was deemed to be inconclusive (for example, the subject has stored some relevant information, but she is not able on the basis of that information to tell whether the sentence in the database was affirmative or negative). In that case, the process would have terminated at the end of the second phase and response times comparable to those for 'Yes' and 'No' answers are to be expected.

In order to test their model further, Glucksberg and McCloskey devised another experiment (experiment 3 in the 1981 paper), in which the two different kinds of don't Know answers were made easier to distinguish. In that experiment, subjects' use of don't Know answers was tested for statements pertaining to general knowledge. They made a distinction between two kinds of questions: questions for which they expected subjects to have relevant information in their memory (such as 'Does Ann Landers have a degree in journalism?') even though that information was bound not to be sufficient to answer the question; questions for which they expected subjects not to be able to find relevant information (such as 'Does Bert Parks have a degree in journalism?').[9] As predicted by the two-stage model, Don't Know responses of the second kind, were found to be matched by faster reaction times.

The main interest of Glucksberg and McCloskey's study from our perspective is that it rests on a division between two forms of don't know answers that can be related to the division between what we called uncertainty and unawareness. On their account, a first category of don't know answers corresponds to cases in which the agent has *no idea* of what the answer might possibly be (because subjects never had access to the answer). Cases of unawareness can be subsumed under that category. Suppose I am asked 'Is Tolkien the author of *The Hobbit*?', and never heard of either Tolkien or *The Hobbit* before. From Glucksberg and McCloskey's model we expect that lack of acquaintance with the concepts will give rise to a fast and reliable 'don't know' answer. Glucksberg

[9] Ann Landers and Bert Parks were two public figures widely known to most Americans in 1981. Ann Landers is the pen name of an advice columnist. Parks was a television announcer and the emcee of the Miss America pageant. What Ann Landers is known for is relevant to the question whether she has a degree in journalism. What Bert Parks is known for is not directly relevant to the question whether he has a degree in journalism.

and McCloskey's second category of don't know answers corresponds to cases in which the agent does have relevant information, but hesitates between competing possibilities. This fits what we characterized as uncertainty proper.

It is important, however, to stress that not all cases that Glucksberg and McCloskey describe as cases for which the agent has no idea of what the answer might be can be adequately described as cases for which the agent has no acquaintance with the conceptual ingredients relevant to the question (the latter is certainly a sufficient condition for lack of any idea about the answer, but not a necessary one). For instance, in a question such as 'Does Margaret Thatcher use an electric toothbrush?', the expectation is indeed to get a don't know answer of the first kind, but not because subjects never heard of Margaret Thatcher or do not have the concept of an electric toothbrush. Rather, the idea is that subjects will simply lack any appropriate evidence on which they can base their answer. In this regard, the dichotomy we propose between uncertainty-related unknowns and unawareness-related unknowns does not exactly coincide with Glucksberg and McCloskey's dichotomy, although it can be subsumed under theirs as a particular case.

Hampton et al.'s 2012 study

In a more recent study, Hampton et al. (2012) have proposed a distinct measure in order to evaluate what people know about their ignorance. Instead of comparing response times attached to don't know answers, they investigated whether the possibility of using an ignorance answer, as opposed to just True or False, increases the consistency of the subjects' use of True and False. In their study, subjects had to complete a questionnaire twice: one group of subjects could use only True/False answers, and another group could use '100% sure it's true', '100% sure it's false', and 'Not sure either way' (henceforth, Unsure). Subjects were invited to fill the questionnaire during a first session and then at a second session 1 or 2 weeks later. In the first experiment, subjects in each condition were presented with three distinct kinds of questions: questions of either general knowledge (viz. 'Texas is the size of Oklahoma'), autobiographical facts (viz. 'I have used a blue notebook'), or about category statements (viz. 'Darts is a sport').

Hampton et al.'s basic finding is a more pronounced use of the 'Unsure' answer for questions of general knowledge, as opposed to questions about categories or autobiographical matters. For autobiographical questions or categorization questions, subjects were not significantly more consistent in their answers when they had the possibility to use the 'Unsure' answer, as opposed to just 'True' and 'False'. The 'Unsure' answer only increased consistency for questions about general knowledge. In a distinct experiment, Hampton et al. replicated the same pattern by comparing answers to questions of general knowledge to answers given about personal aspirations (viz. 'I aspire to be on TV') and questions about moral beliefs ('animal testing is wrong'). There too they found consistency to increase only for questions of general knowledge.

Hampton's explanation for this contrast also relies on Glucksberg and McCloskey's model. For questions about general knowledge, Hampton et al. suggest that a reason for the greater stability of 'don't know' answers might be the more frequent lack of relevant information in memory. For instance, subjects may be more inclined to respond twice 'Unsure' to 'Texas is the size of Oklahoma' because they don't find any relevant information in memory (in order to be 100% confident either way). In contrast to that, the lesser stability of 'Unsure' answers for autobiographical facts or personal aspirations on their view is that 'you will always have some relevant basis in memory on which to base your answer. In this case it is a question of trying to retrieve evidence and argument in favour of the statement being true or not'. For instance, one may be less prone to saying 'I don't know' to 'I aspire to be on TV', because one can find reasons either way, and eventually in a way that will decide one for a True or for a False answer.

Hampton et al.'s study allows us to elaborate on the use of don't know answers. To a question of general knowledge such as 'Texas is the size of Oklahoma', a typical way in which one would issue a stable don't know answer is when one does not even have a clear representation of where Oklahoma is located in the US and of what size it might be. This would correspond to a case in which one is initially unaware of what the size of Oklahoma might be: I never came across that information. Faced with the question, I can therefore move to a stable state of conscious uncertainty, in which I know that I lack the basic evidence to adjudicate the question. By contrast, faced with an autobiographical question such as: 'I have used a blue notebook', I can think of many occasions in which I have used a notebook. I then try to find evidence for whether or not, in at least one of those occasions, the colour of the notebook was blue. In a case like this, intuitively, it is harder to be in a stable state of uncertainty, in particular because I know that as a matter of principle, I have had this information available to me.

Relation between the two don't know answers

The upshot of the two sets of experiments we discussed is that there appears to be two kinds of don't know answers. The first kind includes cases in which we can be fairly confident that we do not know the answer, because no answer even comes to mind (as in 'What is the name of the largest store in Budapest?'). The second kind of don't know answer includes cases in which we are able to articulate the answer in principle, but fail to be confident that it is the correct answer. For such cases, a judgement of uncertainty is less reliable, simply because there is a competition between alternatives, based on available evidence in that case (as in 'did you ever use a blue notebook?').

A lot of cases may be mixed, however. Compare, for instance, the two questions: 'Is Alabama the size of Oklahoma?' and 'Is Texas the size of Oklahoma?'. At the moment I am writing, I would not be able to locate Alabama and Oklahoma on a US map, but would be able to locate Texas. The names 'Alabama' and 'Oklahoma' are familiar, however, just like the name 'Texas', but I happen to know more things about Texas than about the other two states (for instance, I know Texas shares a border with Mexico). Intuitively, the first question is a question for which I lack any potential evidence, since from my perspective, 'Alabama' and 'Oklahoma' are merely names of indistinct US states. Faced with the first question, I would therefore respond 'I don't know' without hesitation. I have *no evidence for a 'yes' as opposed to a 'no'*. In the case of 'Is Texas the size of Oklahoma?', I do have some potentially relevant evidence in my memory such as: 'Texas is a fairly large state in the US'. This is a case in which, though I do not know anything directly about the size of Oklahoma, and cannot remember exactly how big Texas looks like on a map, I would be tempted to make a guess (here a guess in favour of the hypothesis that Texas is not the size of Oklahoma).

An important aspect to this example is that while part of the question concerns an item about which I have hardly any direct evidence, I have at least some potentially relevant evidence coming from the other half of the question. Intuitively, this would be a case in which I am less than 100% sure that the answer is true, and also less than 100% sure that it is false, but also in which I am more than 50% sure that it is true. The Alabama/Oklahoma case, however, is one in which I am distinctively close to 100% sure that I don't know. In the Texas/Oklahoma case, this rather is a situation in which, though I am strictly speaking less than 100% sure either way, I am no longer indifferent between the yes and no answer. The difficulty about such cases is that, because we do have partly relevant evidence, we cannot be sure that we won't do better than chance. Equivalently, we cannot be sure that we don't know the answer, in the weak sense of being able to give the correct answer so as to do better than chance.

In concrete cases, therefore, the distinction between Glucksberg and McCloskey's two kinds of don't know answers will not be pure. On the other hand, attention to pure cases is revealing of the structure of higher-order knowledge. Thus, pure cases of conceptual unawareness are cases that will give rise to a clear perception of our ignorance. Franke and de Jager, for instance, point out that an important feature of unawareness is that it is fragile. This means that, in a situation in which I am unaware of who Tolkien might even be, the very asking of the question 'Is Tolkien the author of *The Hobbit*?' breaks the unawareness, and puts me in a state of uncertainty regarding whether the answer is yes or no (see 'Unconscious ignorance and implicit knowledge' section). By contrast, most situations of uncertainty are not situations that result from antecedent unawareness states. In the case of 'Did you ever use a blue notebook?', I easily find positive evidence for a 'yes' as well as for a 'no'.[10] A report that one does not know will be less stable, then, because in principle, one would report uncertainty, rather than yes or no, only when the yes-evidence and the no-evidence balance each other. In the next section, we propose to focus on such cases: that is, we will set aside cases of ignorance resulting from unawareness, to focus on the appreciation of one's uncertainty in cases in which we do find evidence for competing answers.

Evaluating one's uncertainty

The experiments reported in the previous section suggest the following picture of the relation between first-order knowledge and knowledge about that knowledge: in cases in which one lacks any evidence relevant to adjudicate whether a proposition is true or not, one can be confident that one does not know whether the proposition is true or not. Cases of prior lack of acquaintance with the constituents of the question are cases for which one will typically issue a reliable

[10] D. Spector (pers. comm.) points out an interesting connection between the two kinds of don't know answers that we distinguish here and the distinction originally due to F. Knight (1921) in economic theory between 'risk' and 'uncertainty'. He also invites us to clarify the link between our notion of uncertainty and Knight's notion. Uncertainty in our sense and as used in epistemic logic is a generic notion, compatible with both what Knight calls risk and what he calls uncertainty. In economic theory since Knight, 'risk' is associated to the idea of an uncertainty that can be sufficiently precisely quantified, based on a priori or a posteriori statistical knowledge. 'Uncertainty' on the other hand is an uncertainty for which the agent may lack the resources to make any adequate probabilistic quantification (Knight 1921, III.VII. pp. 47–48). An illustration of 'uncertainty' in that Knightian sense is given by the classic example from Ellsberg (1961), in which you know that there are 90 balls in an urn, 30 red, and 60 blue or yellow. The ignorance of the exact proportion of blue balls is a case of uncertainty. Another example, this time related to unawareness, appears in a paper by Gilboa et al. (2009): 'There is a semi-popular talk at your university, titled, "Cydophines and Abordites". You are curious and may listen to the talk (. . .) however, before the talk you have no idea what the terms mean. (. . .) You are asked whether all cydophines are abordites. Obviously, you have no idea. But if you are Bayesian, you should have probabilistic beliefs about this fact. How would you be able to come up with the probability that all cydophines are abordites?'.

In our terms, what the example illustrate is a case of uncertainty resulting from antecedent unawareness—here lack of conception—corresponding to a transition of the kind we illustrated in Fig. 20.1 (except that, in this case, and unlike in our Tolkien and Hobbit example, there is not even a stable representation of what the expressions might possibly mean). Gilboa et al.'s remark about probabilities in a sense supports our observation that for uncertainty resulting from conceptual unawareness, there is no competition between relevant sources of evidence, hence no obvious way in which an agent could assign probabilities to the alternatives at issue, in contrast to cases of uncertainty resulting from a competition between alternatives informed by memory and by an adequate representation of conceptual space.

don't know judgement. By contrast, in cases in which one does have evidence regarding the status of the proposition, but only partial evidence, knowing whether one knows or not is typically harder to adjudicate, for it depends on an evaluation of the weight of one's evidence. In this section we propose to examine more closely the relation between the strength of one's first-order evidence and the adequacy of metacognitive evaluations of one's own knowledge.

Discrimination tasks

A relevant paradigm for the investigation of the evaluation of one's uncertainty is given by discrimination tasks in which the difficulty in discriminating is gradual and can be modulated. This paradigm, in particular, plays a central role in Smith et al.'s theory of uncertainty monitoring (Smith et al. 2003). Typically, subjects are assigned a task of discrimination in which they have to report a particular condition (*s*, for signal), or its absence (*n*, for no signal). Subjects are allowed to give three kinds of response, either a Signal Response (*S*) or a No Signal response (*N*) or they can opt out and use a third response (*U*, for 'Uncertainty').

For instance, Smith et al's. density discrimination task is one in which subjects are shown a box with a number of pixels illuminated. The signal condition in their display is when the number of pixels is exactly 2950. No signal corresponds to the case of fewer than 2950 pixels. The signal condition is matched with a Dense response, and the No signal condition with a Sparse response. As expected of such psychophysical tasks, what Smith et al. report is that when the number of pixels is sufficiently below 2950, discrimination is easy and subjects make correct use of the Sparse response. Close to 2950 pixels, subjects make larger use of the Dense response. Between those two levels, there is a range of pixel configurations for which the use of the third response gradually increases (in humans and certain species of animals). Typically, the use of the third response is highest where the response curves for Sparse and Dense cross each other, namely are at a ratio of 1:1. What appears, in particular, is that subjects make the heaviest use of the third response in the region where the competition between signal and noise is at its maximum, namely where subjects are equally drawn to the Sparse or the Dense response.

A noteworthy feature of the curves shown by Smith et al. is that when the third response is at its maximum use, it is only used about 70% of the time (ibid., fig. 3, subject D). In contrast to that, when the Sparse response is at its maximum use (for easy Sparse configurations), it represents close to 100% of the responses. One may wonder about the sources of this asymmetry.

It is possible, first of all, to conceive of experimental set-ups in which optimal use of the Uncertainty response would reach 100%. As signal detection theory (SDT) has made clear, how much a response is used in a task is a function not only of the subject's sensitivity, but also of the structure of rewards and penalties attached to the task (see McNicol 1972/2005). Because of that, a subject could be in a state of less than perfect discrimination about the correct response, that is a state of uncertainty, but still be rationally motivated to use the response Sparse instead of the response Uncertain because the former is a more profitable strategy.

Consider however a structure of rewards such that the expected value of the Uncertainty response in certain configurations is the same that the expected value of the Sparse (or Dense, for that matter) response on other configurations. If one observed a tendency for subjects to use the Uncertainty response in the former configurations to a lesser extent than Sparse (or Dense) on the latter configurations, this would provide evidence that it is harder to adequately perceive one's being uncertain when one is uncertain than it is to perceive that one is certain when one is certain. The response curves presented by Smith et al. are compatible with this hypothesis: they indicate that the uncertainty response is less stable for intermediate cases than the Sparse and Dense responses are for clear cases.

Higher-order knowledge and imperfect discrimination

Incidentally, situations of imperfect discrimination have been used by Williamson (1990, 1994, 2000) as typical exemplifications of cases in which agents ought to lack an adequate representation of their knowledge and uncertainty. Williamson's argument is a normative argument about the failure of positive introspection in principle, but it is interesting to try and relate it to actual tasks of discrimination. Williamson does not use the framework of signal detection to give a model of uncertainty, though some links can be made between his model and the SDT model (see Egré and Bonnay 2010). In particular, he does not rely on a probabilistic representation of uncertainty, as in SDT, but instead he uses qualitative models of uncertainty of the sort informally presented earlier, in which agents can be basically in three states: either they know that *p*, or they know that not *p*, or they are uncertain either way.

For Williamson, situations of imperfect discrimination can be characterized as situations in which the relation of epistemic possibility between alternatives is non-transitive.[11] For instance, for configurations with more than 1000 pixels illuminated (and less than 4000, say), an agent with limited discrimination may not be able to reliably discriminate between pixels configurations that differ from each other by fewer than 25 points. This means that if the pixel configuration were 2950 pixels, the agent could not accurately discriminate it from a configuration of 2925 pixels, nor from a discrimination of 2975 pixels. However, the agent may be able to discriminate a configuration with 2925 pixels from one with 2975 pixels.

In Williamson's approach, the minimum difference reliably detectable between two configurations can thus be viewed as setting a margin of error: between 1000 pixels and 4000 pixels, for instance, the agent estimates the status of configurations with a margin of error of about 25 pixels. This means that if the configuration is exactly 2950 pixels, the agent should be uncertain whether it is Sparse or Dense. If the configuration is 2920 pixels, however, the agent is in a position to know that the configuration is Sparse, because it is represented as having at most 2945 pixels, which still counts as Sparse.

An important assumption of Williamson's account is that margins of error constrain first-order knowledge and higher-order knowledge in a uniform way. This means that if the margin of error were 25 pixels, then in order to be certain that the configuration is Sparse, the configuration must be below $(2950 - 25) = 2925$ pixels. But in order to be certain that one is certain that the configuration is Sparse, the configuration must be below $(2950 - 2 \times 25) = 2900$ pixels. And similarly at higher levels: each new iteration of knowledge is represented by the addition of a new margin of error. This model makes the following interesting prediction regarding the relation between first-order discrimination and metacognition. The prediction is: the further away a signal is from the boundary, that is, the higher the signal to noise ratio, the more confident the agent should be that there is signal. On the other hand, the model makes a counterintuitive prediction, which is that given a fixed margin of error, there will always be a failure of higher-order knowledge at some point (for instance, suppose the agent sees a configuration of 2875 pixels: this is three margins of error away from the Dense condition. Here the agent is predicted to know that he knows that the configuration is Sparse, but not to know that he knows that he knows this, a very counterintuitive prediction). Arguably, however, some pixel configurations are such that the

[11] This assumption is common to a number of other frameworks. See also Halpern (2008), Luce (1956), and van Rooij (2010) among others.

signal to noise ratio will be so high that subjects will be in a position not only to issue a correct answer, but also to be certain that it is the correct answer.

One possibility to amend Williamson's basic model in this regard is simply to assume that perceptual margins of error do not constrain first-order knowledge and metacognitive levels in the same way (see Bonnay and Egré (2009), Dokic and Egré (2009), and Egré and Bonnay (2010) for a conceptual and logical elaboration, as well as Loussouarn (2010) for a discussion of the cognitive basis of a such a distinction). If that assumption is relaxed, one can preserve the idea that the stronger the signal to noise ratio will be, the more confident an agent should be that he has first-order knowledge, while making some cases such that the agent is in principle fully confident about those. Some other options have been suggested: Mott (1997), Dutant (2007), Halpern (2008), and more recently Spector (submitted) basically consider that Williamson's margin for error principles are not sound, as a result of which they preserve a similar idea, which is that positive introspection can be maintained for the internal perceptions of the agent.

Conclusion

Our initial observation in this paper has been the idea that in order to examine what it takes to know that one knows or to know that one does not know something, one should carefully distinguish between different forms of ignorance. Based on a central distinction made in formal epistemology, we have argued that there are two fundamentally distinct sources of ignorance, namely ignorance based on uncertainty, and ignorance based on unawareness.

Both uncertainty and unawareness are sources of unknown unknowns and of unknown knowns. In cases of unawareness, unknown unknowns are the most typical cases, whereas unknown knowns correspond to cases of implicit knowledge that an agent cannot represent to herself for lack of the relevant concepts. In cases of uncertainty, unknown unknowns and unknown knowns are better described as forms of overconfidence and underconfidence respectively, that is as cases in which the agent does not have an adequate representation of the structure of his or her first-order uncertainty and evidence.

Based on that, following Glucksberg and McCloskey's two stage model of answer processing, we have argued that the principled distinction between uncertainty and unawareness has a metacognitive correlate. Deciding whether one knows or does not know the answer to a question appears to give rise to stabler and faster verdicts for cases in which one lacks basic acquaintance with the answer, that is for cases in which one can realize one's unawareness prior to the question. By contrast, deciding whether one knows or does not know the answer to a question is typically harder and more demanding for cases in which one has some evidence available in memory, but competing evidence, that is for situations of uncertainty. Fundamentally, the reason why it is harder to know whether one knows or does not know a proposition in cases of uncertainty is due to the fact that one needs to weigh the strength of available evidence. This stage is simply not needed for cases in which no evidence is found in memory.

Little has been said here finally about the psychological processes that enable us to weigh our first-order evidence and to decide about the strength of our first-order uncertainty. In the case of Smith et al.'s discrimination paradigm, however, we pointed out that there likely is an asymmetry between the confidence that one is certain, and the confidence that one is uncertain. This implies, on our view, that high degrees of first-order certainty should tend to go with high confidence that one is certain. By contrast, high degrees of uncertainty do not appear to give rise to high confidence that one is uncertain to the same extent.

Acknowledgements

We are very grateful to Joëlle Proust for detailed comments on the first version of this paper, and for a number of helpful exchanges on the topic of metacognition. We are also indebted to James Hampton for directing our attention to Glucksberg and McCloskey's model of ignorance and to his recent work on metacognition. We also thank David Spector for stimulating discussions on the epistemology of higher-order knowledge, and Mikaël Cozic for valuable comments and discussions on the topic of unawareness.

References

Aloni, M. and Egré, P. (2010). Alternative questions and knowledge attributions. *The Philosophical Quarterly*, 60(238), 1–27.

Aloni, M., Egré P., and de Jager, T. (in press). Knowing whether A or B. *Synthese*.

Bonnay, D. and Egré, P. (2009). Inexact knowledge with introspection. *Journal of Philosophical Logic*, 38(2), 179–228.

Bonnay, D. and Egré, P. (2011). Knowing one's limits. An analysis in centered dynamic epistemic logic. In P. Girard, M. Marion, and O. Roy (Eds.) *Dynamic Formal Epistemology*, pp. 103–25. Berlin: Springer.

Bromberger, S. (1987). What we don't know when we don't know why. Repr. in Bromberger S. *What we know we don't know*. Chicago, IL: University of Chicago Press and CSLI Publications (1992).

Cozic, M. (2011). Probabilistic Unawareness. *Cahiers de recherche, Série Décision, Rationalité, Interaction*. IHPST: Paris.

Dekel, E., Lipman, B. L., and Rustichini, A. (1998). Standard state-space models preclude unawareness. *Econometrica*, 66 (1), 159–73.

Dutant, J. (2007). Inexact knowledge, margin for error and positive introspection. In D. Samet (Ed.) *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge* (TARK XI), pp.118–124. Louvain-la-Neuve: Presses Universitaires de Louvain.

Egré, P. (2008). Reliability, margin for error and self-knowledge. In V. Hendricks and D. Pritchard (Eds.) *New Waves in Epistemology*, pp. 215–50. London: Palgrave MacMillan.

Egré, P. (2011). Epistemic logic. In L. Horsten and R. Pettigrew (Eds.) *Continuum Companion to Philosophical Logic*, pp. 503–42. London: Continuum.

Egré, P. and Bonnay, D. (2010). Vagueness, uncertainty and degrees of clarity. *Synthese*, 174(1), 47–78.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75(4), 643–69.

Fagin R. and Halpern, J. (1988). Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34, 39–76.

Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning about Knowledge*. Cambridge, MA: MIT Press.

Feinberg, T. E. and Roane, D. M. (2005). Delusional misidentification. *Psychiatrics Clinic of North America*, 28, 665–83.

Franke, M. and de Jager, T. (2010). Now that you mention it: Awareness dynamics in discourse and decisions. In A. Benz, C. Ebert, G. Jäger, and R. van Rooij (Eds.) *Language, Games, and Evolution* (Lecture Notes in Computer Science, Volume 6207), pp. 60–91. Berlin: Springer.

de Gelder, B. (2010). Uncanny sight in the blind. *Scientific American*, 302(5), 60–5.

Gilboa, I., Poslewaite, A., and Schmeidler, D. (2009). Is it always rational to satisfy Savage's axioms? *Economics and Philosophy*, 25, 285–96.

Glucksberg, S. and McCloskey, M. (1981). Decisions about Ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory* 7, 311–25.

Halle, M. (1978). Knowledge unlearned and untaught: what speakers know about the sounds of their language. In M. Halle, J. Bresnan, and G. A. Miller (Eds.) *Linguistic theories and psychological reality*, pp. 294–303. Cambridge, MA: MIT Press.

Halpern, J. (2008). Intransitivity and vagueness. *Review of symbolic logic*, 1(4), 530–47.

Hampton, J. A., Aina, B., Mathias Andersson, J., Mirza, H. Z., and Parma, S. (2012). The Rumsfeld Effect: the unknown unknown. *Journal of Experimental Psychology: Learning Memory & Cognition*, 38, 340–55.

Heifetz, A., Meier, M., and Schipper, B. C. (2006). Interactive unawareness. *Journal of Economic Theory*, 130, 78–94.

Hemp, D. (2006). The KK (Knowing that one Knows) Principle. *Internet Encyclopedia of Philosophy* http://www.iep.utm.edu/kk-princ/

Hill, B. (2010). Awareness dynamics. *Journal of Philosophical Logic*, 39(2), 113–37.

Hintikka, J. (1962). *Knowledge and Belief. An Introduction to the Logic of the Two notions*. Ithaca, NY: Cornell University Press.

Hintikka, J. 1970. Knowing that One Knows reviewed. *Synthese*, 21, 141–62.

Knight, F. (1921). *Risk, Uncertainty and Profit*. Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Co.

Lihoreau, F. (2008). Knowledge-how and ability. *Grazer Philosophische Studien*, 77(1), 263–305.

Loussouarn, A. (2010). *De la métaperception à l'agir perceptif*. Thèse: EHESS.

Luce, D. (1956). Semiorders and a theory of utility discrimination. *Econometrica*, 24(2), 178–91.

McNicol, D. (1972/2005). *A Primer of Signal Detection Theory*. Psychology Press, reissued 2005, Mahwah, NJ: Lawrence Erlbaum and Associates.

Mott, P. (1998). Margins for error and the Sorites paradox. *Philosophical Quarterly*, 48(193), 494–504.

Pérez Carballo, A. (submitted). Structuring logical space.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 159(2), 271–95.

Ryle, G. (1971). Knowing how and knowing that. In *Gilbert Ryle: Collected Papers*, Vol. 2, pp. 212–25. New York: Barnes and Nobles.

Schaffer, J. (2007). Knowing the answer. *Philosophy and Phenomenological Research*, 75(2), 383–403.

Smith, J. D., Shields, W. E., and Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26, 317–73.

Spector, D. (submitted). Margin for error semantics and signal perception.

Stalnaker, R. (1990). Mental content and linguistic form. Philosophical Studies 58. Repr. in Stalnaker, R. *Context and Content*, pp. 225–40. Oxford: Oxford University Press (1999).

Stanley, J. and Williamson, T. (2001). Knowing how. *The Journal of Philosophy*, 98(8), 411–44.

van Benthem, J. (2004). What one may come to know. *Analysis*, 64(282), 95–105.

van Ditmarsch, H., Kooi, B., and van der Hoek, W. (2007). *Dynamic Epistemic Logic*. Synthese Library volume 337. Berlin: Springer.

van Rooij, R. (2010). Vagueness and linguistics. In G. Ronzitti (Ed.) *The Vagueness Handbook*, pp. 123–70. Dordrecht: Springer.

Williamson, T. (1990). *Identity and Discrimination*. Oxford: Blackwell.

Williamson, T. (1994). *Vagueness*. London: Routledge.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.

This page intentionally left blank

Index

This page intentionally left blank

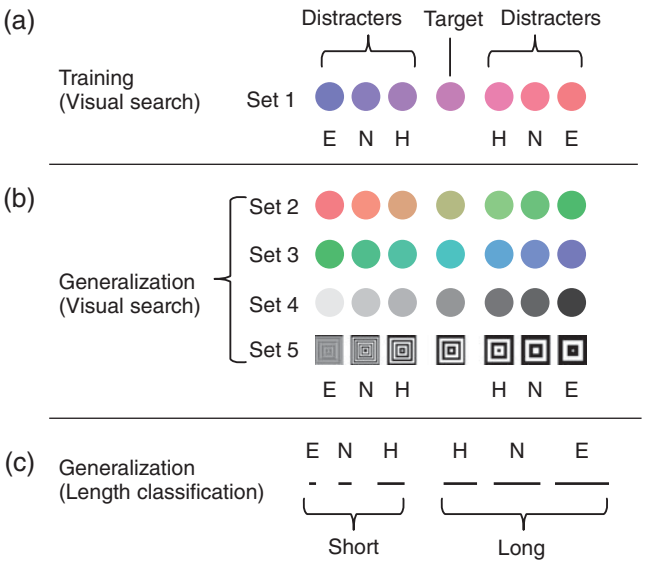


Plate 1 Stimuli used in Study 1. a) Stimuli used in the training of visual search. b) Stimuli used to test generalization to visual search of new items. c) Stimuli used to test generalization to bar-length discrimination. E, N, and H denote easy, normal, and hard discrimination, respectively. With kind permission from Springer Science+Business Media: *Animal Cognition*. Do birds (pigeons and bantams) know how confident they are of their perceptual decisions?, 14(1), 2011, 83–93, Nakamura, N., Watanabe, S., Betsuyaku, T., and Fujita, K. (see also Fig. 3.1)

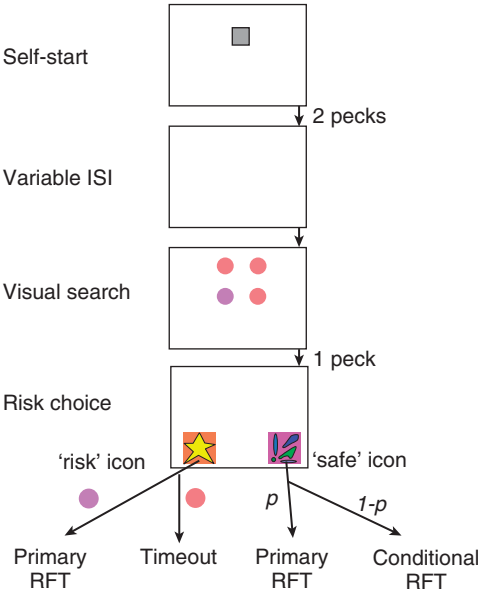


Plate 2 A schematic diagram of the visual search task with risk choice. With kind permission from Springer Science+Business Media: *Animal Cognition*. Do birds (pigeons and bantams) know how confident they are of their perceptual decisions?, 14(1), 2011, 83–93, Nakamura, N., Watanabe, S., Betsuyaku, T., and Fujita, K. (see also Fig. 3.2)

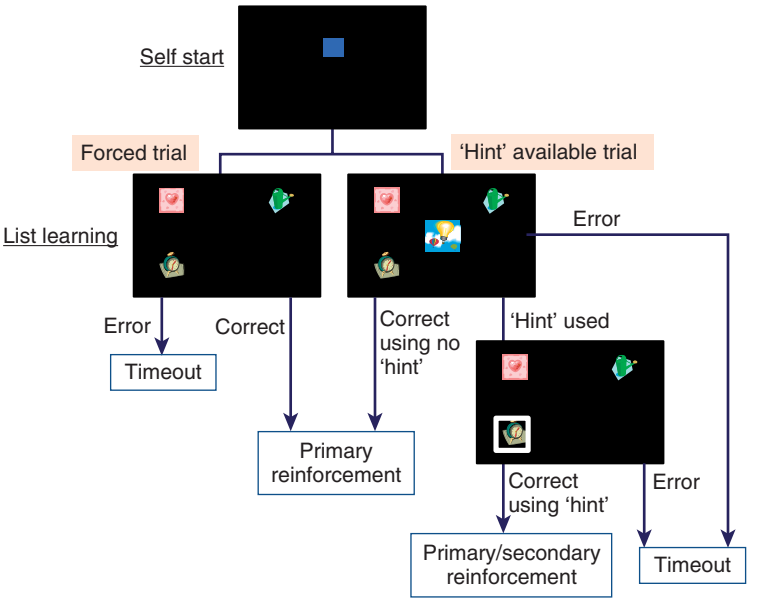


Plate 3 A schematic diagram of the simultaneous chaining task with 'hint' option. (see also Fig. 3.4)

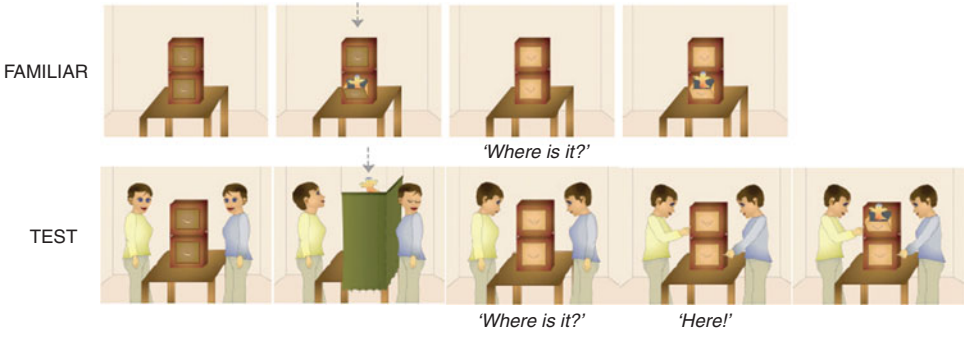


Plate 4 Exemplar stimulus of the eyetracking-task used by Neumann (2009). Reproduced from Neumann, A. (2009). Infants' implicit and explicit knowledge of mental states. Evidence from eye-tracking studies (unpublished dissertation). (see also Fig. 7.1)